

格致方法 · 定量研究系列　吴晓刚　主编

项目功能差异（第二版）

[美] 史蒂文 · J. 奥斯德兰（Steven J. Osterlind）
霍华德 · T. 埃弗森（Howard T. Everson） 著

周韵 译

- 革新研究理念
- 丰富研究工具
- 最权威、最前沿的定量研究方法指南

37

格致出版社　上海人民出版社



格致方法 · 定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）

上架建议：社会研究方法
ISBN 978-7-5432-2206-9



定价: 15.00元
易文网: www.ewen.cc
格致网: www.hibooks.cn

格致方法·定量研究系列　吴晓刚　主编

项目功能差异（第二版）

[美] 史蒂文·J.奥斯德兰（Steven J.Osterlind）
霍华德·T.埃弗森（Howard T.Everson） 著
周　韵 译

SAGE Publications, Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

项目功能差异:第2版/(美)奥斯德兰(Osterlind, S. J.),(美)埃弗森(Everson, H. T.)著;周韵译. —上海:格致出版社:上海人民出版社,2013
(格致方法·定量研究系列)
ISBN 978-7-5432-2206-9

Ⅰ. ①项… Ⅱ. ①奥… ②埃… ③周… Ⅲ. ①社会科学-研究方法 Ⅳ. ①C3

中国版本图书馆CIP数据核字(2012)第299597号

责任编辑 王亚丽

格致方法·定量研究系列
项目功能差异(第二版)
[美]史蒂文·J. 奥斯德兰 霍华德·T. 埃弗森 著
周韵 译

出 版	世纪出版集团 www.ewen.cc	格致出版社 www.hibooks.cn 上海人民出版社

(200001 上海福建中路193号24层)



编辑部热线 021-63914988
市场部热线 021-63914081

发 行	世纪出版集团发行中心
印 刷	浙江临安曙光印务有限公司
开 本	920×1168毫米 1/32
印 张	5.25
字 数	81,000
版 次	2013年2月第1版
印 次	2013年2月第1次印刷

ISBN 978-7-5432-2206-9/C·96
定 价 15.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的 SAGE 出版社定量社会科学研究丛书中的 35 种，翻译成中文，集结成八册，于 2011 年出版。这八册书分别是:《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》、《数据分析方法五种》和《列表数据分析》。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与 SAGE 出版社合作，陆续推出新的品种。我们希望本丛书单行本的出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研

究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院(Institute for Social Research)定量社会研究方法项目的一些讨论会，受益良多。

2003 年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的

方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。“社会科学里的统计学”（Statistics for Social Science）从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种

语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此

我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

我们希望本丛书的出版，能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

序

从广义上说，社会科学中的定量研究有两大任务。首先，社会科学研究者希望能够用特定测量尺度捕捉并描摹社会现象，而这一测量尺度必须对应该社会现象的概念化。其次，研究者通常希望能够用某种解释性统计模型描摹社会现象，以便从模型中更深刻地理解现象的成因。

本系列中绝大部分丛书可被归于测量和建模这两个类别中（其中，建模类在数量上稍占上风）。此外，仅有一小部分丛书是有关统计与数学方面的，或是探讨一个截然不同的研究模式（例如个体为本模型——然而即便对于个体为本建模而言，测量和建模方面的知识也是至关重要的）。

测量的历史源远流长。远古时代，人们为了生存即有了测量的需要，如建造房屋、缝制衣物、交换生存所需的食品货物等。然而，除基本生计外，人们也需要测量：若是没有足够好的测量，我们就不会有斯通亨治巨石阵。然而，直到近代，我们才有了近乎社会科学意义上的测量。约翰·格兰特(John Graunt)于1662年所作的《关于死亡表的自然与政治观察》可被看作测量人类生存的第一次严肃尝试。而对于教

育和心理学研究中如何建构有用的测量，这样的思考则是更近的事情了。直到1904年，爱德华·桑代克(Edward Thorndike)才意识到，在拼写能力测试中使用拼对的词数为标准存在无法避免的模糊性。

无需赘言，对于社会科学研究而言，测量(或教育与教育心理学中的测试)极其重要。我们也需要明白，测量为何对本系列丛书同样至关重要。继奥斯德兰(Osterlind)于1983年出版《测试项目偏差》后，相关研究欣欣向荣。如今，测试项目这一术语本身也不足以描述这一领域中的问题了。也正因为如此，奥斯德兰与埃弗森(Everson)合著了这本新书《项目功能差异》。本书并不仅仅是对奥斯德兰1983年著作的简单延伸——项目功能差异这一术语，其含义就已经超过了测试项目偏差。简言之，项目功能差异指的是，测试项目在经过能力配对后的不同社会群体(如种族、性别和阶级)间存在功能性差别。因此，尽管这一方法主要立足于教育与心理学，却对广义的社会科学研究具有重要意义。

本书是对旧作的及时更新，更为本丛书带来了领域内的新发展。在本书中，奥斯德兰和埃弗森回顾了一系列项目功能差异统计方法，包括传统的Mantel-Haenszel检验、似然比检验，以及logistic回归应用(这一方法在本系列其他丛书中，常被作为一种解释性模型——而不是对测量的评估——加以讨论)。尽管本书着重关注教育与心理学，然而从测量的角度说，其他社会科学领域的读者也将大获裨益。

廖福挺

目 录

第1章

导　言

本书旨在取代奥斯德兰于1983年所著的《项目偏差检验》一书。《项目偏差检验》为本系列(即《社会科学中的量性应用》)的第30本书。在过去的25年里,对于组间在检验项目中存在的可识别的差异表现,这方面的心理测量研究经历了重大的发展,因而仅仅对过往的著述进行更新已远远不够。项目调查的具体技术有所改变(一些旧方法被抛弃,新的技术取而代之),不仅如此,更关键的是,“项目偏差”这一概念本身也被重新界定。与局限于单个项目的早期研究相比,当下的相关研究大多立足于更加宽广的视角。如今,我们不仅着眼于结构表现(construct representation)中的测试公平(test fairness)与组内无差性(group invariance),还关注项目功能差异(Differential Item Functioning, DIF)。此外,一些研究者正试图从实证角度,识别不同组间表现差异的原因(Allalouf、Hambleton & Sireci, 1999; Roussos & Stout, 1996)。在这一背景下,“项目功能差异”并非仅是对“项目偏差”的更确切表述,而是评估中广义上的现象。它涉及心理评价中一系列不同方面,对于测试的开发和检验有着重要的影响。它是项目和测试功能方面尚未有定论的研究的一部分。

立足于"项目功能差异"这一更宽广的视角，本书旨在向读者呈现项目功能差异及相关领域的前沿理论与研究，特别是项目功能差异的识别。通过讨论一系列相关共发问题，进而介绍一系列可用于研究项目功能偏差现象的统计方法。本书希望能够帮助读者广泛了解有关项目功能差异的方法技术。然而本书涉及的话题深广，并非薄薄一本书所能涵盖。因而贯穿本书，我们也将引述其他对进一步研究测量偏差、项目偏差和项目功能检验有助益的资料。这是一个充满活力的广阔领域，相关文献众多，并在不断发展。

第1节 | 读者与背景知识

本书针对学者、测试研发者、学生及其他对测量与项目表现(item performance)有兴趣的读者。由于绝大多数用以研究项目功能差异的策略依托于统计学,因而在统计描述和统计推论方面具有坚实基础至关重要。读者同时需要了解方差分析方法,如ANOVA(方差分析)、ANCOVA(协方差分析)及其多变量对应形式(即MANOVA与MANCOVA)。同时,需要熟悉相关关系及其相应操作。此外,在回归分析及其相应操作,特别是logistic回归方面的知识也十分必要。由于许多信息由公式表述最为准确,因此读者需要能够理解相关公式。本书中的公式,通篇使用约定俗成的标注方式,并且在必要的时候,我们将对公式中的特定项目做出注释。另外,读者需要了解概率论相关元素及其在似然函数中的表述——这些对于若干项目功能差异的测定操作都十分重要:如那些依托于项目反应理论和logistic回归的步骤。最后,我们认为读者在传统测量理论(即真分数理论和经典测量理论;CTT)和项目反应理论(IRT)方面有扎实背景,包括相关技术与假设。我们知道,有越来越多的读者已经掌握了相关知识、技巧和能力。

情境中的项目功能差异及其补救方法

稍后，我们将讨论发现项目功能差异后的一些补救办法。然而，必须从一开始就指出，项目功能差异是一个极其复杂的现象，因而在许多谨慎准备的教育和心理测量中都会有一定程度的存在。由于项目功能差异是一个常见现象，因而并不能因为它的出现，就将一个测试——或者是“测试”本身——认定为不公平或是有偏的。在精神测量中，这一现象过于复杂和不确定，因而无法简单地一概而论。而测量的结果也不应该被草率地抛弃。

此外，若认为将一个检测出项目功能差异的项目从测试中撤去就能够改善整个测试，这样的认识也是彻头彻尾错误的。这样的想法不仅天真，简直荒谬。尽管有时表现出项目功能差异的项目需要被舍弃，但是在大多数情况下，它们可以被修订，甚至不需要做任何改正。我们强调，项目功能差异是一个极其复杂的现象，在特定的测量情境中，我们应当采取一个审慎成熟的态度加以对待。

第2节 | 项目功能差异研究的角色

在对测试项目的广泛调查中,项目功能差异方法的使用及其功能的首要目的十分明确:项目功能差异着眼于确定被试者对于一个或者一系列测试项目的回答,是否与其个体特质(如性别、种族)相关。我们认为,对项目功能差异的调查,有助于我们更彻底地研究一个测试项目,改善其测量方式或是更准确地理解测试分数。然而必须指出的是,项目功能差异并不在于在评估和测试中揭示真正的组间差别。毕竟,估计真正的区别本身是评估的目的。

项目功能差异的调查首要在于评估测试效度。正如先前所指出,项目功能差异的研究在于帮助验证对测试分数的预期解读,并进一步帮助评估基于此分数所做出的结论(Kane,2006)。《教育与心理测量标准》(美国教育研究协会、美国心理学协会、全国教育测量委员会,1999)有力地指出,效度是量表研发和使用的首要考量。使用包括项目功能差异分析的多方论据,对于有意义地、恰当地、有用地解读测试分数十分必要。因此,项目功能差异方面的证据对效度的帮助体现在许多方面:包括量表的研发和评估、在决策框架下使用测试分数,例如教育、就业项目的选拔或对候选人的认证。

第3节 | 公平性与项目功能差异

在当下对项目功能差异的理解中，公平性问题至关重要。在此，我们简单探讨项目功能差异和公平性的精髓，以便后文详细讨论。在考虑测试结果是如何被用于进行决策时，我们最容易理解测量公平性的概念。显而易见，依据测试结果进行决策是认知评估的要义，而要想理解精神测量中的项目功能差异，必须要充分理解测量公平性。然而，公平地进行依托于测试的决策，却是核心要旨。当然，一个充分认识到特定情境和可能后果的决策同样也是智慧的。因此，公平性不仅是对所涉对象的敏感，也与人类智慧相关。

当然，公平性也存在于法律语境中。在美国，其起源是宪法的第十四条修正案。通过于1868年的第十四条修正案指出，“人人都有平等地获得法律保护的权利”。这一修正案进一步强化了“人生而平等”的理念，这一宣言对项目功能差异研究有重要而直接的影响。更近一些，1964年的民权运动再次强调了公平性的概念，如今这一概念也被不断地再次强化。法学家指出了民权法律在测量公平性方面的广泛应用。例如，卡米利(Camilli，2006)就曾对测量方面的公平性立法进行了编年研究。

第 4 节 | 偏差与项目功能差异

当基于测试分数的决策不公平,或对某一群体产生不同的影响时,测试偏差便出现了。例如,标准化测试的批评者通常将男女生在数学分数上的差异当做存在测试偏差的证据。项目偏差将这一概念由测试延伸至项目层面,例如,某一个测试刺激不公正地反映了某个群体或其信仰。几乎在所有情况下,测试或项目偏差将导致不公平。这一概念进一步衍生了克利里于1968 年提出的著名的、测试中统计偏差的更狭义的概念:

> 当测试对总体中某一组测试者存在偏差,即在测试旨在评估的那一个方面,对于某一组测试者,始终存在不为 0 的估计误差。换言之,即根据回归分析估计的分数曲线,对于某一组成员始终是过高或者过低的(Cleary, 1968:115)。

克利里于 1968 年提出的对于不公平预测的定义,多年来都十分有用。然而,在对测试区别的现代的广义的理解中,它略显囿限。这一概念更接近于我们如今对某些识别项目功能差异的技巧的描述。克利里成果的首要贡献在于,她将我们的注意力转向了测量中这一重要的方面,因而她的影响也留存至今。

第5节 | 项目功能差异与歧视

尽管在通俗语境下,“歧视”(discrimination)通常意味着某种程度上的不公平,常与偏向、不宽容甚至狭隘仇视相关,但在测试的心理测量评估中,它有完全不同的含义。在心理测量研究中,这一词语指一个测试及其组成项目的区分度。在测量中,它通常被分类为“项目区分度”。事实上,在心理测量研究中,区分度是测试的积极方面。信度,是测试区分度一致性的指标之一。想象一把无法区分出一英寸与其他长度的尺子:显而易见,一个没有区分度的测量是无用的。因而,从测试区分度的角度出发,充分理解所有项目功能差异识别技巧,将有用的测试区分度与其余产生的与目标建构无关的组间区别区分开来,是至关重要的。

为了系统地进行上述区分,在项目功能差异研究中的不同组必须具有可比性。通常,许多项目功能差异方法将每组进一步分为组内组,以便细化配对。总而言之,在所有项目功能差异识别技巧中,在评估指标上根据能力进行组间配对,都是至关重要的一方面。

第6节 | 理解“项目功能差异”之概念

至此,可见在研究项目表现中的组间差别时,三个概念需要被同时考量:测量公平性、测量偏差和项目功能差异。作为一个概念,项目功能差异统计学包含了一系列对于发掘测试和项目层面上系统的组间差别有用的统计方法。研究项目功能差异没有单一的技巧方法,相反,项目功能差异作为一个心理测量名词,描述了一系列对于测量的一个具体却重要的方面十分有用的统计方法。此外,项目功能差异并不指示所识别差异的方向,也不指示可能的因果关系。最后,项目功能差异的应用被限制在统计和心理测量研究中。组间差异的政策与政治讨论,需要同时考虑三个相互关联的方面:公平性、测量和项目偏差、项目功能差异。

第7节 | 本书使用语言

本书的首要目标，也是与1983年版的最大区别之一，就是旨在用清晰的方式，加深对项目的组间测试差别这一领域的了解。为了阅读便利，测试与精神测量两个概念互换使用。此外，我们常使用“项目”代指“系统研发的测试刺激”这一更广泛的概念。当然，测试刺激有多重形式，例如多项选择、评分、其他的类别性回答、不同的结构性回答，甚至表演等。项目的具体形式（例如二分类或多分类评分项目）在许多项目功能差异识别技巧中，是被预先假定的，也可以通过我们的表述看出，因此，我们并不屡次重复具体形式的名称。当然，当需要准确描述时，我们也将做出标注。

此外，尽管并非完全同义，但我们将“能力”（ability）与“技能”（proficiency）两词互换使用。“能力”暗指隐性的、被评估的结构（appraised construct），而“技能”则更倾向于指某种技巧和心理发展的才能。然而，尽管在许多心理学语境下这样的区分相当重要，但在项目功能差异方面，是否区分两者并不存在实质区别。

最后，在讨论教育和心理测试时，我们仅指那些根据特定标准研发的、结果具有信度与效度的标准化测试。总的来说，它们依据《教育与心理测量标准》（美国教育研究协会等，1999）中的标准。

第 8 节 | 本书所用示例

为了所涉及不同项目功能差异技巧的解释具有一致性，我们使用一个测试，即大学基本学科考试(College BASE)的不同项目进行阐释。这是一个评价大学生在英语、数学、科学和社会科学方面知识、技能及一定推理能力的成就性测试。其核心在于检测学生在大学通识教育大纲下特定技能与推理能力的掌握程度(Osterlind、Sheng、Wang、Beaujean & Nagel，2008)。

当然，由于不包括任何识别性信息，我们的统计数据不会泄露特定的测试和项目。此外，我们所用的所有示例来源于已经不再使用的项目。一个考试的不同项目使得我们的示例具有一致性，因而不同的方法可以被合理地对比比较。

最后，本书所涉及项目功能差异统计量由一系列软件计算而成，其中包括主流软件如 SPSS 与 SAS，以及一些专门软件如 BILOG-MG、MULTILOG、PARSCALE 和 SIBTEST，这些都在后文中标注了出来。

第2章

项目功能差异描述

眼下,各类教育和心理测试在美国和世界范围内广泛使用。在许多社会经济领域,大学入学考试、入职测试、精神健康调查及许多其他心理和教育评估被用来为政策和实践提供信息。测试使用者通常认为,测试分数在不同组别之间是具有可比性的,根据这些分数也可以得出公平的比较。然而,如果生成和使用的测试分数或测试项目偏向某一个组别,那么根据测试所获得的推论的效度,将受到质疑(Kane, 2006; Messick, 1989、1988)。当有证据表示某一个测试不公平地偏向特定的人口组别(例如男人和女人,黑人与白人),则通常认为这样的测试是具有偏向性的。在这些情况下,通常怀疑测试项目在不同群体间有功能性的区别。测量专家将这样的区别称为"项目功能差异",简称 DIF(Dorans & Holland, 1993; Holland & Thayer, 1988; Holland & Wainer, 1993)。

从效度的角度讲,项目功能差异值得关注,其原因在于,当项目功能差异出现时,通常意味着不同的人口群体"在根据心理特点或其他特质进行配对后,能否在测试项目上获得成功仍有不同的可能性"(Clauser & Mazor, 1998:31)。正如其他研究者所指出的,承认在能力配对后仍可能存在表现

差别，这一点非常关键，因为这意味着仅仅观测到测试分数中存在差别，并不能构成测试偏向性的证据。当培训和学习的机会分配不公时，不同人口组别的成员有时会在能力上有所区别。在这种情况下，其结果通常被界定为“项目影响”而非项目偏向(Camilli，2006；Camilli & Shepard，1994；Clauser & Mazor，1998)。

第1节 项目功能差异正式定义

在心理测量的文献中,人们对"影响"与"测量偏差"或项目功能差异做出了区分(Camilli, 2006; Camilli & Shepard, 1994; Dorans & Holland, 1993; Millsap & Everson, 1993)。"影响"指的是组间差异在测试和测试项目表现中的体现。例如,在教育与应聘测试中,个体和组在测试所测量的特点上往往有区别——因而存在"影响"。例如,在普遍使用的数学测试(例如SAT,即学习能力测试;NAEP,即全国教育进步测试)中,男孩往往比女孩得分高(Wilder & Powell, 1999)。

与测试的"影响"不同,项目功能差异或测量偏差指的是,测试项目在根据能力配对后的不同组别中,仍存在功能性的差异(Camilli, 2006; Camilli & Shepard, 1994; Holland & Wainer, 1993; Penfield & Camilli, 2009; Zumbo, 1999)。正如我们和其他学者(见 Holland & Wainer, 1993)所强调的,当评估项目功能差异或测量偏差时,为避免辛普森悖论,必须评估配对组在表现上的区别(Clifford, 1982; Simpson, 1951)。在这个悖论下,在配对的个体间,项目影响的方向和组间区别的方向是不一致的。如果仍用数学举例的话,比如,一个数学问题对于女生总的来说是更难的,然而,对于测

试者中一部分经过能力配对的人而言，它对于这其中的女生来说，却是更简单的。我们往往在大学入学考试的测试题中看到这样的例子。

因此，为了解释明晰，我们在总的层面对测量偏差给出正式定义，以便我们更好地理解什么是项目功能差异。依据米尔萨普（Millsap）和埃弗森于 1993 年的研究和彭菲尔德（Penfield）与卡米利在 2007 年的研究，我们用一个简单明晰的例子解释这一概念：Y 代表对某一测试项目的回答，θ 代表被测试的潜在结构（或能力），根据规范，Y 是 θ 可观察的指示项。在这一框架下，我们可以将 Y 在 θ 上的概率分布用函数 $f(Y)\mid\theta$ 表示。假设我们需要考察对于两组测试者（焦点组与对照组），Y 的有条件概率分布。从统计角度说，无论哪组被当成焦点组或对照组都应该是一样的。然而，在文献中，我们通常认为对照组由测试偏向的个体组成（例如主流群体，或是男性），而焦点组则由可能在测试中不占优势的个体构成（例如少数群体，或是女性）。因此，给定函数 $f(Y)\mid\theta$，并进一步假设焦点组和对照组的测量误差分布相同，我们可得：

$$f(Y \mid \theta,\ G = R) = f(Y \mid \theta,\ G = F) \qquad [2.1]$$

此处，G 代表分组变量，R 代表对照组，F 代表焦点组。公式 2.1 即是不存在项目功能差异的情况。

需要指明的是，Y 的概率分布与组别无关，从而进一步证明了不存在项目功能差异。例如，假设存在一个二元的测试项目，$Y = 1$ 代表回答正确，$Y = 0$ 代表回答错误，在这种情况下，在 θ 值上相等的个体，无论属于哪个组别，回答正确

的可能性都是相同的。显然,这意味着该测试项目不存在项目功能差异,因为无论是焦点组的成员还是对照组的成员,只要具有相当的特点和能力,回答正确的可能性都是同样(或者说,近似同样)的。因此,不存在一个组比另一个组更占优势的情况。

这一定义与其他定义的区别,在于对"影响"和"项目功能差异"的区分。公式 2.1 中无法观测的有条件不变性(unobserved conditional invariance)也是洛德(Lord)在 1980 年对项目偏差缺失的定义的一部分——这一定义,常常被认为是对此概念的最基本的描述。至今,洛德的定义都是项目反应理论应用中项目功能差异检测的基础(Thissen、Steinberg & Wainer, 1988)。我们再次强调,在项目偏差定义中包含"控制 θ",对于区分测量偏差或项目功能差异与普通的组间区别和项目影响十分重要。例如,不同组可能在 Y 分数分布上存在差异,或用公式表达:

$$P(Y \mid G = g) \neq p(Y) \qquad [2.2]$$

然而,普遍认为,公式 2.2 并不足以证明以上所定义的偏差存在(Ackerman, 1992; Drasgow, 1987; Holland & Thayer, 1988; Lord, 1980; Millsap & Everson, 1993)。测试者的表现与 θ 有关,而不同组别的 θ 不同,公式 2.2 可能在偏差不存在的情况下仍显示偏差。在实证研究中,偏差的调查通常假定不同组别极有可能在 θ 上具有差别。

现在,我们进入逻辑上的下一步,假定对于焦点组和对照组而言,Y 的有条件分布存在不同。也就是说,不同组别具有相同能力和特点的测试者,Y 的概率分布是不同的。在

这种情况下,即便控制 θ,仍存在组别和项目表现的相关关系。当一个测试项目是二元时,这样的有条件相关(conditional dependence)意味着焦点组和对照组具有相同 θ 的测试者,回答正确的概率是不同的——对于这个测试项目,回答正确概率较低的组别,是不占优势的。因此,我们认为测试项目在这两组间的功能表现不同,即存在项目功能差异。

由此可见,在一定程度上,所有项目功能差异检验统计都试图证明"不存在项目功能差异"这一原假设,或试图测量在多大程度上,公式 2.1 是无法确证的。

第2节 | 项目功能差异:统一与否

读者一定不会奇怪,对项目功能差异的研究已经困扰测试研发者们一段时间了。教育测试机构(ETS)研究测试偏差的先驱威廉·安格夫曾说:"测试研发者们常说,他们常常面对项目功能差异而无法理解,再多思考似乎都也无法解释,为何一些完全合理的项目却具有极大的项目功能差异值。"(William Angoff, 1993:19)

测试研发者们所面对的模棱两可和困惑往往在于,人们通常假设,测试是一个一维的测量方式。的确,项目功能差异的存在意味着测试测量了一些次级因素(例如考试速度,或者不同测试者的考试技巧,例如猜题)。然而,在更多时候,我们往往不知道项目功能差异的最深层原因。

然而,无论如何,为研发公平无偏的测试,测试研发者通常需要展示是否有次级因素存在,如果存在,它们是否是测试设计特点的一部分。正如我们从项目功能差异研究中所学到的那样,在测试明细中所需要(或暗示)的次级因素,其分布对于焦点组和对照组而言往往是不同的。因此,为了充分研究所观察到的、往往是难以解释的项目功能差异是否是由于某些次级因素,测试研发者往往需要依赖专家对测试项目内容的分析。例如,如果项目功能差异可以被归因于某些

无关的测试项目内容，或是其他不需要的项目特点（例如进行测试的方式，项目在测试中的位置等），这样的项目便可以被视为是不公平的。

而让问题进一步复杂化的是，存在两类项目功能差异：统一的和不统一的。如我们所见，项目功能差异是在测试表现和组别的相关性框架下进行讨论的。统一的项目功能差异是最简单的一种，在此，对于不同值的 θ 而言，测试表现和组别的相关关系都保持一致。也就是说，在整个 θ 分布中，对照组在这一具有功能差异的项目上的优势是一贯的。换言之，存在统一的项目功能差异意味着，对于特定组别（对照组），其成员，无论能力如何，在该项目上成功的概率，始终是高于相应能力的焦点组成员的（Mallenbergh, 1992）。当然，在一定情况下，这样的相关关系根据测试者潜在能力不同而有所不同。不统一的项目功能差异为，当 θ 值不同时，这样的相关关系方向与大小也发生变化。正如测试者常常发现的那样，对于某个项目而言，某个 θ 级别的对照组测试者可能具有小到可以忽视的优势，而在另一个较高 θ 级别的对照组测试者则具有较大的优势。更让人困惑的是，测试者有时也会遇到"交叉"的项目功能差异，即在 θ 值的一个区域内，对照组具有相对优势，而在 θ 值的另一个区域内，焦点组具有相对优势。如图 2.1 所示，统一和不统一的项目功能差异可以通过对于两个项目的曲线表现出来。

如图 2.1 所示，项目特征曲线（Item Characteristics Curve）提供了一个方式，可针对两个不同组别（如焦点组和对照组）对于同一个项目的应答进行比较。项目特征曲线中的差异意味着，具有同样能力的、属于不同组别的测试者在同一个项目

上回答正确的概率是不同的。两个组别的曲线始终不同且不存在交叉,即意味着统一的项目功能差异。两个组别的曲线不同,并在一定 θ 值时交叉,则意味着不统一的项目功能差异。而两条曲线中的面积,意味着项目功能差异的程度。

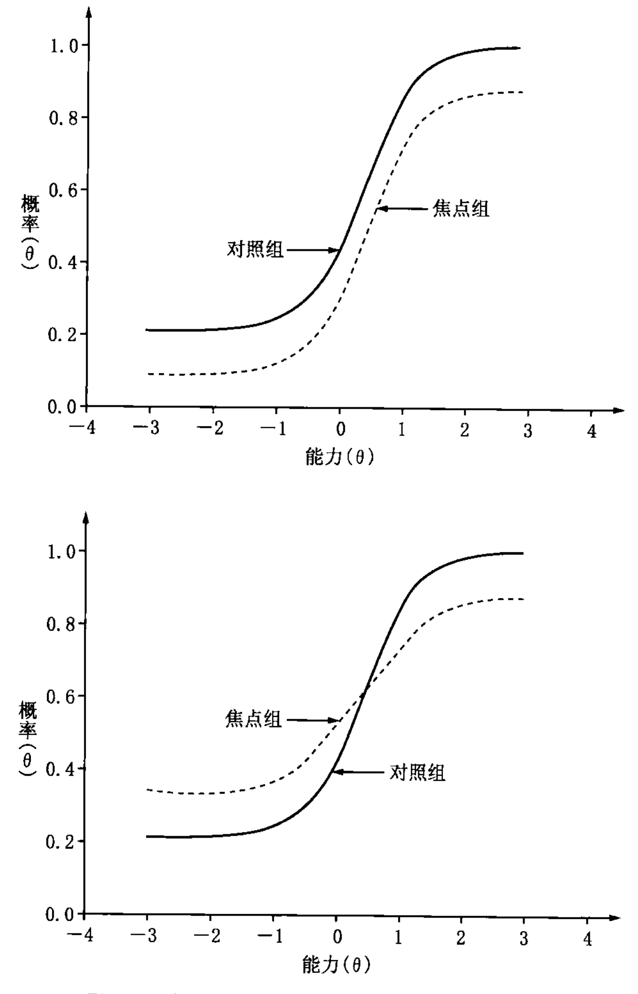


图 2.1 统一(上图)和不统一(下图)的项目功能差异

如果两条项目特征曲线存在交叉，那么一部分区域可被视为存在正向项目功能差异，另一部分则被认为存在负向项目功能差异(Camilli & Shepard, 1994)。对于理解与阐释项目功能差异而言，这样的区别十分重要。而根据差异是否统一，我们可能需要使用不同的统计检验方法。

第3章

项目功能差异的统计侧面

第1节｜并非均值区别

必须要注意的是，测试人群中不同组别间平均分数(中位值或均值)存在差别，并不能自动地被归因为项目功能差异，甚至不能证明存在项目功能差异。简单的组间分数均值差异，其本身也不反映项目功能差异的任何重要方面。毕竟，精神评估(即测试)的重要一点是可信地发掘出某个测试项目中能力的区别，以及组间区别。当然，分析组间分数的差异是一个分析组间区别的常用统计方法。因此，项目功能差异并不是可观测的分数差异的假设。

与这一点紧密相关的是，在项目功能差异研究中，组间差异的原因与被测量的结构无关。也就是说，项目功能差异之所以存在，在于测试目标建构以外的原因影响了不同组测试者的表现。例如，如果在一项数学测试项目中，发现在男孩女孩两组间存在功能差异，这意味着性别与测试的目标测试建构(数学能力)无关。但是，十分重要的是，我们并不知道为什么不同性别在测试表现中会有所不同。我们仅仅发现，在两个组别间，具有数学意义上显著的、系统性区别。当然，数学上的显著并不等于统计意义上的显著，某些项目功能差异检验并不能证明其统计意义上的方差区别。

第2节 | 测量误差

当检验项目和测试中的项目功能差异时，必须要理解测量误差的概念。测量“误差”的概念与日常语境中的“错误”有本质的区别。在日常生活中，我们说“犯了错误”，暗含的意思往往是，这样的错误，如果更加谨慎是可以避免的。在测量科学中，对于误差的定义是更加技术性与复杂的。它与计算中的方差有关。教育与心理学测量中的方差来源于概率论与统计学，它指的是一个随机变量的统计分散。从数学意义上，方差的定义是标准差的平方。在一个正确抽样的样本里，误差存在，是由于一个给定值与期望值(即均值)不同。在样本中，聚合起来即为方差。一个给定值在方差上的偏离即为误差。类似地，信度即在一个特定的测量方法中发现误差，它由一系列系数组成。项目功能差异从本质上说也是一个测量误差。此外，测量误差被分为两类：随机误差与系统误差。奥斯德兰将其区别描述如下：

> 认识到存在多种误差——或者更确切地说，误差来源——十分有用。总的来说，误差来源被分为随机的和系统的两类。随机误差是给定个体真实分数与观测分数的区别，而系统误差则包括与测试设定或所测量的能

> 力无关的、一致的组间区别。区分这两类测量误差在于其目标。与给定个体相关的是随机误差,与组间反应相关的是系统误差(Osterlind, 2006:59)。

项目功能差异方法在很大程度上被用于挖掘后者(系统误差)的来源。然而,在此领域 30 多年的研究后,我们发现,项目功能差异的来源并非如我们所想的那样,是系统的。同时,它们也并不容易被轻易发觉。例如,有时候我们只在焦点组的一部分中发现存在项目功能差异。项目功能差异的来源在很多时候是非常难以察觉的。

第3节 系统性的区别

正如之前已经提到的那样，在检测项目功能差异时，另一个重要的需要考量到的方面在于，组间分数区别在某种程度上必须是系统性的。也就是说，组间区别必须用近乎相同的、可预测的方式影响所有人。最常见的情况是，在回答一个测试项目时，焦点组与参照组相比，是不占优势的。这一概念(项目功能差异)描述了测试和项目影响。据此，项目功能差异描述的不是一个个体在回答测试项目时的表现。它描述的总是组间现象。

第4节 | 能力配对

据此，显而易见，检测项目功能差异中重要的一步，即不同组在所测试能力方面的配对。这样的配对通常是根据全部测试分数实现的。在绝大多数项目功能差异方法中，测试总分被视为衡量测试者能力的标志。据此，能力相似的测试者在某一个测试项目上的应答可被进行比较。

同时，被比较的小组对于更大的测试群体(或者说测试者的总体)也至关重要。也就是说，在项目功能差异的检测中，焦点组和对照组的选择必须和某个适当的目标相关。例如，我们通常会关注不同性别和种族间的项目功能差异。这些组别和心理测量等其他许多方面的相关研究十分相关。然而，如果要研究"左撇子"和"右撇子"或者有不同头颅大小的人群间的项目功能差异，这些组别就毫无用处。

第5节 | 控制条件:外部依据与内部依据

正如我们已经强调的,对于检测项目功能差异至关重要的一点,是不同组别的个体必须具有相当的能力。因此,如何确定"相当"能力即成为了关键的一点。

首先需要考量的是,配对依据对于项目功能差异检测的测试而言,应当是外部的还是内部的。一个外部的依据可以是另一个和目标测试的评估目标相关的测试和评估。例如,对于一个测定学校和知识水平的测试而言,智商测试可以被看做一个外部的依据。

在能力配对时,外部依据之所以有用,在于它不太可能被目标测试中与项目功能差异无关的因素(例如测试时间,测试运行方法)影响。然而,使用外部依据也有相应的风险,需要遵循一系列重要原则:首先,使用的外部依据必须具有一定的信度和效度。其次,与内部依据一样,外部依据必须适用于特定测试目的。例如,在能力测试的项目功能差异检测中,性格测试便不能作为外部依据。所使用的项目功能差异的检测方法也会影响外部依据的选择。汉布尔顿(Hambleton)、博沃德(Bollward)及同事于1990年对外部依据与内部依据的比较研究后发现,对于Mantel-Haenszel(M-H)方法,其结果是

相似的。然而,他们的研究在广度上有一定局限性。因此,由于此领域研究的局限,研究者在使用外部依据进行配对时,必须谨慎。

在绝大多数情况下,一个恰当的外部依据可能并不存在,因而,只能采用内部依据。在此,我们的做法是采取测试总分作为配对依据。测试总分最能反映项目功能差异测试的特点,因而是一个合理的配对标准。然而,在这种情况下,必须充分考虑测试本身的信度和效度。最重要的是,研究者必须充分了解配对依据。对于谨慎的项目功能差异研究,仅仅是因为可以使用总分而使用总分作为配对标准,往往不够谨慎。

第6节 | 数据清洁

在使用内部依据时，一个需要考虑的问题是，在依据中，是否应该包括一个可能具有功能差异的测试项目。清洁指的是将一个可能具有项目功能差异的项目从配对依据中移除。是否对测试进行清洁，这一选择并不容易，在很多情况下，在并没有发现存在项目功能差异时，移除一个项目往往是不合理的做法。然而，当我们怀疑某个测试项目具有功能差异时，移除它往往是合理的做法。一个解决办法是，进行初步的项目功能差异检测，然后移除怀疑的项目，再次计算总分。清洁的一个优点在于，操作容易，并且可以与任何项目功能差异统计方法，包括依于项目反应理论的方法，同时使用。克劳泽(Clauser)和梅热(Mazor)在 1998 年的研究中表示，这一方法并不会提高一类错误的出现率。1995 年，斯托特(Stout)将此整合进 SIBTEST(同步项目偏差检测)方法的一部分。SIBTEST 方法，是一个使用非参数统计方法的项目功能差异检测方法(见 Shealy & Stout，1993)。当然，清洁也可以从一个更复杂严谨的角度对待。

两个更复杂的清洁方法是：迭代清洁与二步清洁。这两种方法都试图解决统计检验力和假阳性的问题。这一方法，首先由伯恩鲍姆(Birnbaum)于 1968 年在洛德和诺维克

(Novick)的《精神测试分数的统计理论》一书中提出。之后，谢泼德、卡米利、埃夫丽尔(Shepard、Camilli & Averil, 1981)与希利和斯托特(Shealy & Stout, 1993)的研究，进一步丰富了迭代清洁方面的研究。在迭代清洁中，首先确认测试项目中显著水平最高的项目值。通常，我们使用 M-H 项目功能差异方法进行这一步。然后，我们排除具有功能差异的项目，再次计算统计值。如果原先的被判断有功能差异的项目仍是显著的，我们再移除显著水平第二高的项目，并再次计算统计值。这一步骤迭代进行，直到我们达到了预先设定的显著水平，或是预先设定的测试次数。此外，除了移除一个显著水平最高的项目，我们也可以同时移除显著水平最高的两个项目。这样，迭代可以更快地完成。

二步清洁的方法最先由霍兰(Holland)和塞耶(Thayer)于 1988 年提出。第一步，项目功能差异的统计值(通常使用 M-H 方法)被计算出来，当获得显著水平的临界值后(无论采用 α 为 0.05 或是 0.01)，该项目被移除。第二步，使用同样方法，再次计算测试统计值。研究者(见 Kwak、Davenport & Davison, 1998)指出在这时检验力较强，假阳性出现地更少。

第4章

重要考量

在项目功能差异中,有一系列重要考量。我们在此讨论其中的一些,研究者需要了解这些问题,注意它们可以更好地帮助阐释与理解项目功能差异的结果。

第1节 项目功能差异:统计判断与经验判断

在研究精神测量中的项目功能差异时,研究者首先需要注意的一点是,分析测试项目,不仅需要心理测量与统计学方法、经验判断,智慧也是十分必要的。即便在分析看似简单明了时,例如在检验一个项目的难度与区分度时,这样的智慧与技巧也是十分必要的。而在检测群体间不同组别间的项目功能差异时,这样的专家智慧和经验就更加必要了。这是由于,理解项目功能差异的结果通常不是简单明了的。通常,在使用不同的检验方法时,我们可以获得不同的检验结果。因此,研究者需要特别注意,检测项目功能差异不仅需要充分使用统计和测量方面的专业知识,同时也需要运用常识。用明智的方法对待项目功能差异检验,毫无疑问可以帮助我们更好地理解这一领域的研究。

研究者发现,对于某一类精神测试和项目,项目功能差异通常存在,意识到这一点尤其必要。成熟的研究者知道,这并不意味着测试通常是偏向性的、不公平的。项目功能差异检测本身是一种有着各种假设和不确定性的统计方法。正如一切统计方法一样,假阳性可能出现。尽管正确解读其结果是有用的,但项目功能差异的研究是一项不精确的科学。

第2节 | 统计偏向并非不公平

在进行项目功能差异研究时,必须意识到,心理测量意义上的偏差并非统计意义上的偏差。在统计研究中,偏差是期望值与真实值间的差异。在统计中,无偏估计对于研究总体化是非常有用的。然而在项目功能差异研究中,偏差指的是对于相当能力的测试者,对特定项目的应答中存在的系统的差别。

第3节 | 完整测试与单个项目

当研究者开始研究项目功能差异时，一系列操作问题应运而生。一个常被提出的问题是，在研究项目功能差异时，究竟是应该关注整个测试，还是单独研究具有差异的项目。我们建议研究者将测试中的所有项目作为一个整体研究。因为，项目功能差异并不是单纯阅读测试项目就能够发现的。因此，在分析之前就将特定项目分离出来研究，这样的做法令人怀疑。再者，测试者将测试项目放在一起，以有效地反映所测试的问题。根据测试构架的不同，以及缩放比例等其他心理测量方面的原因，不同测试中需要进行功能差异分析的项目数也是不同的。一般而言，具有20个项目的测试，就足够进行项目功能差异分析了。

另一个常被提出的问题是，进行项目功能差异分析需要多少测试者。当根据种族确定焦点组和对照组时，常会出现不同组别人数不同的情况，因而研究者需要特别注意各样本的相对大小。当然，研究者需要注意相关组别的大小与分布，及其正态、线性以及齐差性情况。这里的重点问题是样本是否合理，而这并非本书重点关注的话题。然而，需要注意的是，项目功能差异研究中的假设涉及的是两个独立组(焦点组与对照组)。对于样本相互独立性的测量，可以帮助

我们确定有关项目功能差异研究所需要的样本大小。独立组 t 测试可以帮助回答这些问题。此外,当正态分布与差异性假设无法满足时,可以采用例如曼—惠特尼(Mann-Whitney)测试的 U 统计量之类的分参数测试,帮助我们进一步了解这一问题。

第4节 | 数字:项目与样本

在项目功能差异研究中,有三个数值需要考虑:第一,焦点组与对照组中的测试者人数;第二,测试中进行功能差异检验的项目数;第三,测试总长度。对于这三者中的每一个数值而言,都有一系列需要测试者考量的问题。这里的关键问题是,大小必须足够使所使用的统计方法给出有效的结论,同时能够进行一定程度上的总体化。

第三个问题是最简明的,所以我们首先探讨。朱姆伯(Zumbo)于1999年建议,测试长度必须包括至少20个项目。他认为,在将测试总分作为配对依据时,测试总分必须是有意义的。显而易见,缺少一定效度的测试并不能满足这一点。因而,研究者不仅需要审视所评估的构架(construct),同时,需要评估测试的内部特点。在这里,信度和效度是问题的关键。当测试评估的是一个较窄的构架时,通常,得出一个有效的分数所需的项目数也较少。奥斯德兰于2006年进行的研究,是测试评估方面较好的讨论。

第二个问题在于测试中进行项目功能差异检验的项目数。这一数字需要根据不同的测试决定,因而没有经验数值。在有些而并不是所有情况下,将所有项目都包含进去是可行的。正确的方式是,研究者需要首先确定项目功能差异

研究的原因和范围。据此,再对项目数做出有根据的判断,并且确定哪些项目是符合研究目标的。

最后一个问题是焦点组与对照组中的测试者人数。在多数研究中,理论考量和实践操作中的考虑是交叉的。显然,我们知道不同组别间测试者的人数。当分组是根据性别或种族进行时,这一数据是现成的——然而,对于有意义的统计数据而言,却未必是足够的。并且,在有些时候,不同组别的大小可能严重不成比例。有些时候,可以将少数族裔合并为一组,归为"主流"和"非主流"两组。此外,研究者需要特别注意,焦点组与对照组不应严重不成比例。

然而,对样本大小的考量并未就此结束。不同的统计方法往往对样本大小有自己的要求。例如,一个使用卡方的统计方法会比依托于项目反应理论的方法需要更小的样本。此外,样本分布是同样重要的一个方面。许多统计方法都假设了正态分布、线性关系与齐差性。另一些统计方法还有自己的条件,例如,建立在回归分析基础上的方法通常需要检验样本中的异常值和高影响值的情况。确定样本大小可进一步参阅威廉 1978 年出版的《抽样概论》(*A Sampler of Sampling*)。

第5节 | 理性视角

最后需要指出的是，需从一个理性的视角看待项目功能差异。通常情况下，特别是同时考察一个测试中的许多项目时，往往会发现一些项目具有功能差异。然而，当研究者仔细检查具有功能差异的项目时，往往又找不到原因解释这种差异。例如，一个简单的数学题可能会在男孩与女孩间出现功能差异。这时，研究者必须应用自己的技巧、经验与常识。有些时候，出于某种原因，使用检测项目功能差异的方法时会检测出差异，而差异实际上又是不存在的。同时，测试者必须注意，仅仅是一些项目存在功能差异，并不意味着整个测试都要被否定——否则，古往今来所有测试都应该被抛弃了，而根据其他心理测量评价与常识，这显然是不可能的。我们生活在一个不完美的世界里。

第5章

项目偏差与项目功能差异的历史

项目功能差异与测试偏差虽然是一个量性的统计学问题,然而,却应该被置于现当代对于测试公平与机会公平的社会与法学研究背景下加以理解。在本章中,我们讨论一系列相关的重要历史、社会与法律议题。

直到20世纪60年代,测试公平与项目偏差的问题才引起研究者的注意。在美国,约翰逊总统于1964年签署了民权法令。这项法令也标志着美国民权时代的开始。而法令的签署,也让研究者们开始重视在教育与职业测试中可能出现的负面影响。"负面影响",作为一个法律术语,指的是支持非法歧视的表面证据。伯克(Berk)于20世纪80年代早期编著了一套使用新统计方法识别项目偏差的论文集,他指出:"在20世纪60年代后期与70年代早期,心理测量专家加紧步伐,试图为客观标准下的偏差给出定义,并用严谨准确的方法分析偏差,并思考项目偏差的实证研究。"

相似地,南希·科尔(Nancy Cole),教育测试机构前主席,曾在探讨民权运动对于测试公平性的影响时说:"对测试与项目偏差的考量来源于这个时代,被这个时代影响,也是对这个时代的回应。正因为民权运动,对项目和测试偏差的考量,成为了测试基本操作的一部分。"

因此，正是在这样的时代背景下，对测试公平性的研究成了测试研发和使用者所关注的问题。

在后民权运动时代，测试者通过进行公平性测评和研发实证方法，来发掘测试中可能对特定组别不公平的测试项目。而如今，对于测试公平性的测评是绝大多数测试研发者和机构在测试设计和开发过程中重点关注的问题之一(Zeiky, 2006)。

测试公平性的标准

正是由于意识到了心理和教育测试在现代公民社会中日益重要的作用,三大重要组织(美国教育研究协会、美国心理学学会和美国教育测量委员会)共同制定了最新版的《教育与心理测量标准》(美国教育研究协会等,1999)(以下简称“1999 年版标准”)。这些共同制定的标准随时间推移而不断演变,而最新版是这套标准的第四个版本,建基于这三大机构早在 20 世纪 50 年代——民权运动前——的研究。因此,1999 年版标准体现了这一领域各机构在测试公平性问题上在解决社会问题和技术问题两个层面所做的尝试。

在评价 1999 年版标准的重要性时,卡米利对测试公平性的定义做出了如下引用:

> 测试公平性指的是将测试分数用来评估被测者能力时采用的视角。测试公平性概念与测试效度紧密相关。对测试公平性的评估需要一系列广泛证据,其中就包括实证数据等。然而,很多时候,它也包含了法律、民族、政治、哲学和经济方面的考量(Camilli, 2006:225)。

在探讨公平性时,1999 年版标准(美国心理学学会等,

1999)指出,普遍公认的是,测试应该不具有偏差。而对偏差的定义是:

> 当测试本身存在缺陷时,容易出现偏差,从而使得不同组别的被试者的测试分数含义不同。当这种缺陷表现在项目反应层面上时,通常我们使用普遍公认的术语“项目偏差”和“项目功能偏差”。对偏差的考量在测试操作中至关重要(American Psychological Association et al. , 1999:74)。

在描述项目功能差异中的 Mantel-Haenszel 步骤时,霍兰和塞耶指出了相似的一点:

> 可以通过比较特定测试项目、比较总分或比较某个特定评价依据上获得成功的可能性,对两个不同组别做出比较。焦点组,有时被称为被保护组,是关注的重点。它被与另一个组,即对照组比较。之前使用的术语“主流”与“非主流”已不再使用(Holland & Thayer, 1988:130)。

因此,在一个标准的项目功能差异分析中,第一步是定义焦点组与对照组。然后,便可以通过统计方法估计两组间的差别。无论从法律的角度还是测量的角度讲,都必须注意,组间差别可能源于真正的组间教育或心理上的区别。

正如1999年版标准指出的,测试公平性并不意味着相同的测试分数或分类结果。在此,对测试公平性的评估必须通过实证的方式确定,所观测到的组间区别,究竟是由测试范围外的因素所导致,还是由测试架构本身导致。1999年版

标准对检测项目层面上的组间表现区别中的统计偏差,给出如下定义:

> *项目功能差异指的是,相同能力的不同组别的测试者在回答特定项目时出现差异*(American Psychological Association et al., 1999:81)。

对1999年版标准的详细讨论,超出了本书的范围。若要进一步了解1999年版标准,以及它是如何与测试公平性与测试偏差联系起来的,可参阅卡米利于2006年的研究。

回顾过去,可以发现,通过统计方法发掘可能存在偏差的项目,这一研究真正始于20世纪70年代末。然而,近十年后,到了80年代中晚期,一个可操作的统计框架才真正出现,成为在更广泛层面上分析项目偏差的基础。这一框架,即"项目功能差异",由教育测试机构的保罗·霍兰及同事,通过一系列论文首先提出(见 Holland, 1985; Holland & Thayer, 1988)。在之后的二十年中,项目功能差异领域出现了一系列数量繁多的对项目偏差的识别与研究。项目功能差异的方法被整合进测试效度的研究中,并被整合进了1999年版标准(标准7.3)。此外,正如彭菲尔德和卡米利于2007年指出的,项目功能差异已经逐渐进入了心理学与健康科学中其他依赖于标准化测量的领域(见 Bolt, 2002; Dodeen & Johanson, 2003; Gelin、Carelton、Smith & Zumbo, 2004; Lange、Thalbourne、Houran & Lester, 2002)。在这些教育测试之外的领域中,项目功能差异帮助研究者进一步思考在评估组间差异时的效度问题。

第6章

快速而不全面的方法

一些情况下，在正式的项目功能差异检测前，获取一些测试项目表现上组间差异的信息十分有用。一个简单的、初步的观察可以为之后更详细的项目功能差异打下基础。这些初步的简单检测只使用简单的统计学知识与研究方法。而这，恰恰也是它的优势所在。我们将这些方法称为“快速而不全面的方法”，因为正如这个名字，它们提供了有用的、粗糙的指示，但是并不能提供足够准确的信息。因此，在一开始，我们就提醒研究者，这些简单的观察并不能为真正的区别下准确的结论，而简单的统计所获得的结果，也不能作为项目功能差异存在的直接证据。然而，即便如此，这样的初步研究也可以为进一步研究组间差异提供有用的信息。此外，这些快速而不全面的方法的另一个缺点在于，如果组间差别在所有能力层次上不是统一的，这样的方法便检测不出其中的差别。用项目功能差异的术语来说，这些方法只能针对统一的项目功能差异，而不能为非统一项目功能差异提供信息。

第1节｜项目排序法

这样快速而不全面的方法之一，即为比较两组（对照组与焦点组）怀疑有功能差异的项目的排序。在不考虑能力区分的情况下，分别计算两组项目的 p 值。然后，对这些 p 值进行排序与比较，以便检测组间差异。具体步骤见表6.1。

表6.1包含了五个项目的通过率（即 p 值），为了方便比较，它们根据不同组别被排序。如表6.1所示，对于对照组和焦点组而言，项目2具有最高的 p 值，因而意味着较小的组间差异。而项目4的 p 值则表现出了组间差异。对于对照组而言，项目4的 p 值位居第二，而对于焦点组而言，却排在第五。由此初步可见，此项目值得进一步研究。此处需要注意的是，对于这一方法而言，焦点组和对照组没有进行能力配对。因而仅凭借这一信息，并不能做出项目功能差异的解读。然而，尽管如此，这一方法仍然能够快速简便地进行组间对比。

表6.1　两组项目难度的排列次序

项目	对照组		焦点组	
	排序	p 值	排序	p 值
1	3rd	0.64	2nd	0.62
2	1st	0.93	1st	0.81
3	4th	0.55	3rd	0.51
4	2nd	0.71	5th	0.19
5	5th	0.37	4th	0.28

第 2 节 | 能力组法

能力组法对于项目排序法的一个改进在于，将各组根据能力进行区分，并在每个能力层面进行比较。当然，这一方法仍然不能确切地对项目功能差异进行阐释。许多包括 TESTFACT 在内的经典项目分析程序可以进行相关的简便操作(Wood et al.，2003)。在此，我们使用另一程序 MERMAC 进行示范。在此程序中，各组的表现分别独立计算，然后，将各组的能力分布进行分层(在此例中，我们将其分为五层)，见图 6.1。上图为对照组的表现，下图为焦点组的表现。括号中(D)为每一个项目答案的正确答案数原数据，错误选择与 p 值(标记为 DIFF)及点二列相关系数(RPBIS)为给定的。

Reference Group

Matrix of Responses by Fifths for Question 91

D Is Correct Response

	A	B	C	(D)	E	OMIT
5th	0	8	15	135	0	0
4th	2	14	23	114	0	0
3rd	1	8	26	130	0	0
2nd	1	11	32	107	0	1
1st	5	24	37	93	0	5
DIFF	0.01	0.08	0.17	(0.73)	0.00	0.01
RPBI	-0.09	-0.11	-0.15	(0.24)	0.00	-0.15

Percent of Correct Response by Fifths for Question 91

```
5th +                                          *
4th +                                     *
3rd +                                       *
2nd +                                   *
1st +                             *
    +----+----+----+----+----+----+----+----+----+----+
    0   10   20   30   40   50   60   70   80   90  100
```

Focal Group

Matrix of Responses by Fifths for Question 91

D Is Correct Response

	A	B	C	(D)	E	OMIT
5th	4	116	22	10	0	0
4th	6	107	28	11	0	0
3rd	5	110	26	11	0	0
2nd	7	121	26	4	0	0
1st	16	106	25	8	0	0
DIFF	0.05	0.73	0.17	(0.06)	0.00	0.00
RPBI	-0.12	0.05	0.00	(0.02)	0.00	0.00

Percent of Correct Response by Fifths for Question 91

```
5th +    *
4th +     *
3rd +     *
2nd +*
1st +   *
    +----+----+----+----+----+----+----+----+----+----+
    0   10   20   30   40   50   60   70   80   90  100
```

图 6.1　两组在同一项目上根据能力级别进行的经典项目分析

由图6.1可知，两组项目表现从总体上说是存在差异的：对照组与焦点组的 p 值分别为0.73与0.06。此外，可以看出，这样的差别在不同能力层间几乎是统一的。这一方法的优势在于，可以看出组间差别具体在哪儿——在本例中，焦点组较多地选择了错误选项B——因此，研究者可以进一步研究这一项目，试图发掘焦点组倾向选择某一特定错误选项的原因。此外，研究者需要注意的是，尽管组的分层被用做能力分组的替代，然而，这一分层却是独立进行的，并没有进行能力配对，因而构成了项目功能差异调查中的一个很大局限。

为了避免因为比较未经配对的组别而可能产生的问题，必须首先计算配对，然后再进行项目分析。对于严谨的项目功能差异研究而言，在比较之前进行配对，是十分必要的条件。而这样的比较步骤也更接近于M-H方法，我们将在下一章进一步讨论。

第3节 | 过时的 ANOVA 方法

之所以在这里仍然提及通过比较方差比值(variance ratio)的 ANOVA 方法,主要是因为这一方法仍然在文献中被大量涉及,而人们却并没有意识到它的过时之处。尽管早期的研究(如 Berk, 1982; Jensen, 1980)认为简单的 ANOVA 可以被当做有效的检测策略,然而,这一方法如今已被弃用——这一方法仅仅分析组间方差比值,而没有进行项目功能差异检测必需的能力配对。主要通过 SIBTEST 讨论 ANOVA 方法的缺陷,这一方法我们将在后文进一步讨论。然而,一些近期的研究(Whitmore & Schumacker, 1999)通过将 ANOVA 方法和建立在项目反应理论基础上的策略结合,取得了一定的成功。然而,在重新推广这一方法前,还需要更多研究。

第7章

Mantel-Haenszel 步骤

正如我们所见,在早期的项目功能差异研究中,研究者通过 ANOVA 及其他分布统计,试图寻找比较焦点组和对照组的方式。尽管这些方法从一开始就充满问题,最终也并不成功。另一个显而易见的方法是通过采用卡方检验显著性。然而,由于比例并不能在这一检验中直接使用,因此一个简单的卡方检验并不合理。此外,在低频次情况下,卡方测试并不可靠,而低频次恰恰是项目功能差异研究中常常出现的情况。最后,在具有多重自由度的项目功能差异研究中使用卡方检验,它缺少足够的统计检验力来评估项目功能差异的原假设(见 Holland & Thayer, 1988)。因此,我们并不推荐普通的卡方检验方法。

然而,两位 19 世纪 50 年代的医学研究者,生物统计学家内森·曼特尔和流行病学家威廉·亨塞尔,在充分意识到卡方检验的问题后,发展出一套适用于分层样本的卡方方法(Mantel & Haenszel, 1959)。他们试图利用卡方检验可作为测量交叉计算中行列间线性关系这一优势——这对展示组间比较十分有用。曼特尔和亨塞尔的研究是对科克伦(Cochran)统计针对连续性和方差膨胀的小样本修正。二十多年后,霍兰(1985)使用他们修正后的卡方检验,并将此作

为检测项目功能差异的方法。正是由于霍兰(1985)自己以及霍兰和塞耶(1988)的研究,这一项目功能差异研究方法被命名为Mantel-Haenszel(M-H)步骤。如今,这可能是最常用的项目功能差异检测方法。

在本章中,我们着重讨论二分类项目(答案为对或错的项目)中M-H步骤的应用。当然,M-H步骤在多分类项目中也有使用,我们将在之后的章节中展开讨论。

第 1 节 | 卡方列联表

M-H 步骤依据卡方分布，然而使用的是被称为完全卡方(full chi-square)的常见卡方(与列联表)的一个变种。在这里，和四项相关的全息项目因素分析类似，所关注的变量(即测试项目)，从理论上被看做一个从"无"到"完全"的心理连续。经过统计调整后，即可被用来进行项目功能差异分析。

从操作步骤上说，在 M-H 步骤中，对照组和焦点组根据测试总分，被分为不同能力的层次，因而满足项目功能差异分析中所需的能力配对的条件。通常，我们将组分为四到五层，对每一层做 2×2 的卡方列联表。如表 7.1 所示，表中包括了每组每个层次正确和错误答案的频次。需要注意的是，表 7.1 中的数据仅仅包括了一个能力层次，同时其他能力层

表 7.1　单能力层级上单项目的 M-H 卡方表

第 23 项，能力级 2

	正确	错误	总和
对照组(R)	164	97	261
焦点组(F)	102	144	246
总和	266	241	507

次也有平行的卡方列联表，而形成 $2\times2\times k$ 的卡方列联表，其中 k 代表组数。

M-H 的数据整理见表 7.2。表 7.2 所示数据的统假设与组别属性无关。然而，这一假设需要表中所观察的数据满足抽样模型。因此，列联表的边际值是（N_{rj} 和 N_{Fj}）总体参数，其中 R 代表对照组，F 代表焦点组。而观测值是 n_{Rj} 和 r_{Fj} 的随机样本，如表格 a_i，b_i，c_i 和 d_i 所示。

表 7.2 单能力层级上单项目的 M-H 卡方总表

第 i 项，能力级 j

	正确	错误	总和	
对照组(R)	a_i	b_i	261	$N_{Ri}=a_i+c_i$
焦点组(F)	d_i	c_i	246	$N_{Fi}=b_i+d_i$
总和	N_{1i}	N_{0i}	507	
	$N_{1i}=a_i+c_i$	$N_{0i}=b_i+d_i$		

第2节 | M-H 比值比

M-H 计算的第一步为计算 p/q 的比值比。p 代表对于某一个测试项目回答正确的概率,q 则为 $1-p$。在 M-H 步骤中,比值比通常由 α_{MH} 表示,它代表着列联表中行与列间的线性关系。不同作者使用不同的公式计算比值比,而如果使用如图 6.1 的列联表,比值比可以用如下公式计算。

$$\alpha_i = \frac{p_{ri}/q_{ri}}{p_{fi}/q_{fi}} = \frac{\dfrac{a_i/(a_i+b_i)}{b_i/(a_i+b_i)}}{\dfrac{c_i/(c_i+d_i)}{d_i/(c_i+d_i)}} = \frac{a_i/b_i}{c_i/d_i} = \frac{a_i d_i}{b_i c_i} \quad [7.1]$$

其中 p_{ri} 代表对照组(r)在分数区间 i 中回答正确的比例。q_{ri} 代表对照组(r)在分数区间 i 中回答错误的比例。焦点组同理。

如果组间没有区别,则比值比为 1(即 $\alpha_i=1$),这意味着焦点组与对照组平衡,可被理解为不存在项目功能差异。然而,当 $\alpha_i>1$ 时,则对照组在项目上的表现显著优于焦点组。相反,当 $\alpha_i<1$ 时,焦点组在同等能力层的表现显著优于对照组。

更准确地说,α_i 是估计对照组中的个体获得正确答案的概率超过相应的焦点组中的个体时加权平均值的系数。其本质,是一个自由度为 1、p 值小于 0.05 的 t 测试分布。

此外，α_i 也被用于代表总体估计。为了更准确，通常表示为 $\hat{\alpha}_{MH}$，普通比值比。如公式 7.2 所示，这一统计量可通过所有层次配对的组别估计。

$$\hat{\alpha}_{MH} = \frac{\sum_i p_{ri} q_{fi} N_{ri} \frac{N_{fi}}{N_i}}{\sum_i q_{ri} p_{ri} N_{ri} \frac{N_{fi}}{N_i}} = \frac{\sum_i \frac{a_i d_i}{N_i}}{\sum_i \frac{b_i c_i}{N_i}} \qquad [7.2]$$

尽管这一数值代表着特定项目功能差异的总体，但 $\hat{\alpha}_{MH}$ 的数值却非常难以阐释。因而通常被变换为对数形式，如公式 7.3 所示。所获得的指数由 $MH_{D\text{-}DIF}$ 表示。

$$MH_{D\text{-}DIF} = -2.35\ln(\hat{\alpha}_{MH}) \qquad [7.3]$$

这一新形式代表着指数的中心接近。因此，$MH_{D\text{-}DIF}=0.0$ 意味着不存在任何项目功能差异。并注意公式中的负号，这意味着当 $MH_{D\text{-}DIF}$ 为正时，项目对焦点组更有利，而当数值为负时，则对对照组更有利。

图 7.1 中包含了通常被报告的 M-H 统计量。这是 SPSS 16 版的输出表格。当然，包括 STATA，SAS，R 和 Systat 在内的许多其他软件也可以计算 M-H 统计量。

$\hat{\alpha}_{MH}$ 的真实值和对数值都是给定的。在这个例子中，测试者的人数相对较多，然而，一个相对较小的样本同样是具有可比性的。如图 7.1 所示，项目 25 具有性别间的项目功能差异，而项目 26 则不具有（男性为对照组，女性为焦点组）。这是通过与假设值 0 的显著差别表现出来的。

因而，由此例可知，M-H 是一个相对直接的、可以用于许多测试分析的方法。

Crosstab for Gender by Q25				
Count	Gender			
		male	female	Total
rq25	0	115	132	247
	1	145	105	250
	Total	260	237	497

Crosstab for Gender by Q26				
Count	Gender			
		male	female	Total
rq26	0	108	90	198
	1	152	147	299
	Total	260	237	497

Tests of Conditional Independence					
	Chi-Squared		df	Asymp. Sig. (2-sided)	
	Q25	Q26		Q25	Q26
Cochran's	6.520	0.657	1	0.011	0.418
Mantel-Haenszel	6.057	0.516	1	0.014	0.473

Mantel-Haenszel Common Odds Ratio Estimate				
			Q25	Q26
Estimate			0.631	1.161
ln(Estimate)			−0.461	0.149
Std. Error of ln(Estimate)			0.181	0.184
Asymp. Sig. (2-sided)			0.011	0.418
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	0.443	0.810
		Upper Bound	0.899	1.664
	ln(Common Odds Ratio)	Lower Bound	−0.815	−0.211
		Upper Bound	−0.106	0.509
The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption, as is the natural log of the estimate.				

图 7.1 M-H 的 SPSS 结果



非参数方法

不依赖于总体(如概率分布)假设的项目功能差异方法,为研究项目异常功能提供了全新的视角。这些非参数方法不依赖于任何统计或概率模型,因而可以跨测试进行。在这里,项目功能差异被看做测量中的一个问题,而不仅仅局限于一个特定测试。项目功能差异的非参数研究存在若干方法。

第1节 使用 SIBTEST 的项目功能差异

随着项目功能差异研究越来越多地进入测试使用与决策中测试分数适用性的领域，人们也相应地越来越关注功能差异、偏差与不公平的问题。希利和斯托特于1993年的研究，将这一关注具象化为“通过 SUBTEST 进行维度检验”。这一项目功能差异的方法也可以通过一个线性回归的方式改正 M-H 未标准化的问题。希利和斯托特将 SIB 定义为“同时性项目偏差”。作为一个“同时性”方法，它可以检测项目是否具有功能偏差，并测试是否是不公平的。在这一方法中，通过使用一个回归修正，调整了给定特定项目的不同组均值，以修正不同能力测试者在焦点组与对照组中分布不同可能带来的影响。比起 M-H 步骤和其他简单使用测试原始分的方法而言，这一方法能更精确地进行焦点组与对照组的配对，从而降低一类错误的出现率。此外，SIBTEST 在一个无特定模型的环境里检验项目功能差异，因而相比其他方法，SIBTEST 可以更加不受具体测试环境的影响。虽然这从某种意义上是 SIBTEST 作为非参数检验方法的一个局限，然而在项目功能差异研究中却是一个优势。

SIBTEST 项目功能差异方法探讨测试时间项目识别的

标准化。在此所关注的是，对于项目现象而言，定义一个统一的标准和条件。这表现在，这一方法更加关注测试中测量维度的问题，而非仅仅识别一组项目的非正常功能。相应地，SIBTEST 依托于认知理论。换言之，当一个给定项目的一维性被违反时，即可被视为存在项目功能差异。也就是说，SIBTEST 可以独立于项目影响，而分析项目功能差异。

在这一框架下，组间配对是根据一个隐性因变量确定的，而项目功能差异假设多维度性——因为后者恰恰是存在功能差异的表面证据。被测试的结构占据欧几里得空间中的一个维度，而具有功能差异的项目则至少再占据一个额外维度(即次级维度)。这一方法在理论和操作上的优势在于：

> SIBTEST 分析让研究者能够识别项目功能差异何时存在，之后，研究者根据分析之前就已经做出的对效度的假设，决定识别的项目功能差异或偏差是否构成项目功能差异(Shealy & Stout, 1993:159)。

SIBTEST 的这一特点，使其在测试研发阶段对于测试者和心理测量专家而言格外有用。此外，这种在评估早期的考量也可以为测试使用者和政策制定者在分析如何理解测试分数时提供有效的信息。

SIBTEST 的另一优势在于其效率：许多项目可以被同时评估，这对于较长的测试格外有用(见 Gierl、Gotzmann & Boughton, 2004)。如今，SIBTEST 被广泛使用，然而，研究者(见 Fildago、Ferreres & Muniz, 2005; Finch, 2005)指出，许多 SIBTEST 研究使用的是模拟数据，而非被测真正的表

现。无论如何，这一方法对于降低项目功能差异中一类错误的出现率十分有用(Klockars & Lee，2008)。

最后是 SIBTEST 分析通常包括的描述焦点组和对照组差异的图表。图 8.1 即为一个示例，该图同时还包括了对比步骤(即 Dorans 标准化的描述)。

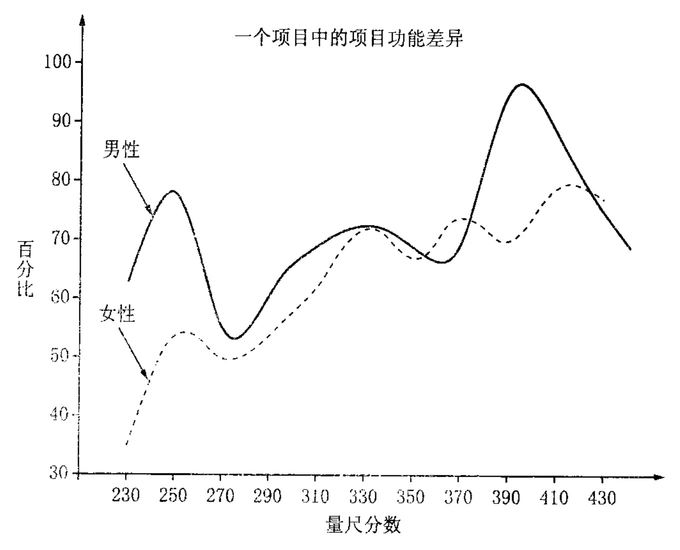


图 8.1 项目的标准化步骤

第2节 | Dorans 标准化

多兰斯(Dorans)和库利克(Kulick)于1983年指出,项目功能差异中的标准问题可作为M-H方法外的另一途径。这被称为"标准化方法"或"Dorans标准化",这一方法旨在使项目功能差异分析的结果更易理解,而M-H方法则更关注于统计检验力。通过首先在大规模的全国性测试SAT测试中使用这一方法,库利克和多兰斯于1986年展示了这一方法的优势所在。在之后的许多年中,这一方法都是项目功能差异研究中的主要方法之一。

正如许多项目功能差异方法和M-H方法一样,Dorans标准化的第一步是根据能力将焦点组与测试组进行配对。这两种项目功能差异方法都可以使用任何合理的标准进行配对,然而通常使用的配对依据是测试总分。如我们在M-H方法中已经描述的那样,这两种方法都将数据纳入2×2列联表中以备分析。和M-H方法一样,在Dorans标准化中,对照组被视为数据的基线(即每一个能力层次所预期的项目表现)。在这一步之后,Dorans标准化出现了和M-H方法的区别。在从列联表中获取组间表现的信息时,与M-H方法使用比值比不同,Dorans标准化分析的是给定项目正确率间的区别。当焦点组与测试组间的p值区别显著时,则被视为

存在项目功能差异。对于每一个项目而言，都有正确率回归于总分量表的散点图，进一步从视觉上呈现项目功能差异。

对于标准化法，图表对于理解项目功能差异有重要的作用。图8.1呈现了一个项目的项目功能差异（即用来描述M-H方法的项目）。在此图中，我们比较了不同分数值上，两组间获得正确答案的被测者的比例。这种数据归纳方式，比简单的多组比较更加完备。事实上，图8.1所呈现的性别和项目反应的关系，比M-H方法中所呈现的更加复杂。正如前文所讨论的，M-H步骤中的组别根据比值比归入不同类别，因此M-H可能会掩盖一定的组间区别，因为每个组在分类时缺乏进一步考量。

第9章

依托于项目反应理论的方法

本章讨论依托于项目反应理论发展而成的项目功能差异方法(见 Hambleton & Swaminathan, 1985; Hambleton、Swaminathan & Rogers, 1991; Lord & Novick, 1968; Osterlind, 2006)。在本章中,我们并不讨论项目反应理论的所有方面,我们关注的重点是项目反应理论中对于项目功能差异研究有用的那一部分,例如项目参数的功能、如何使用项目特征曲线(ICC)和项目反应函数(IRF)等。这一简略的介绍对于在项目反应理论框架下理解项目功能差异十分必要。本章分三小节。首先,我们简要介绍项目反应理论及其在现代教育与心理测量中的意义,特别介绍如何使用项目反应理论方法识别项目功能差异。其次,我们着重介绍如何使用项目反应理论方法识别二元测试项目中的功能差异。最后,我们概述在识别多元测试项目的功能差异中,项目反应理论模型的一系列最新发展。

显而易见,对项目反应理论的完整讨论超出了本书的范畴。然而有相当丰富的资源可以帮助理解这一问题:例如,恩柏里特逊(Embretson)和赖泽(Reise)于 2000 年为教育者与社会科学家提供了项目反应理论的概述;汉布尔顿和思瓦内森(Hambleton & Swaminathan, 1985)与哈布尔顿

(Hableton)及其合作者(1991)写作了两本非常容易理解的著作,是对本话题的概述和介绍;另一本经典著作是洛德于1980年出版的《测试问题中的项目反应理论使用》。此外,配套国际科学软件所发行的项目反应理论软件的使用说明也十分有用:《SSI的项目反应理论:BILOG-MG, MULTILOG, PARSCALE. TESTFACT》(du Toit, 2003)。

在过去20年中,学者们研究出了一系列依托于项目反应理论检验项目功能差异的方法。通过检验单个测试项目或者一组测试项目上的组间差别,项目反应理论方法对于检验项目功能差异,无论从理论层面还是实践操作层面,都十分有用。与经典测试理论相比,依托于项目反应理论的方法对于项目功能差异现象的研究更加全面。在心理测量,特别是教育测试的文献中,记载了大量使用项目反应理论检测项目功能差异的案例(见 Clauser & Mazor, 1998; Millsap & Everson, 1993; Osterlind, 2006; Penfield & Camilli, 2007)。

第1节 | 项目反应理论的框架

项目反应理论关注教育和心理测试中被测者的隐性特点(如能力、态度等),以及用于反应它们的项目的特点。简单而言,项目反应理论探讨测量中的两个基本方面:估计测量刺激的特征和推断被测者的隐性能力。项目反应理论来源于一系列用以缩放测试和推测被测者隐性能力的数学方法。尽管项目反应理论背后的数学知识十分高深,但它们在一系列测量和评估的语境下都十分有用。需要再次强调的是,在项目反应理论的框架下,对每一个测试项目特征的估计和计算,是独立于被测者的能力的。这与传统的测试理论相比,是一个显著优势。项目特征可以和被测者能力在同样的连续轴上表现,对于检测项目功能差异而言,无疑是很有用的。

在有关检验有偏差的测试项目的统计方法的较为完整的早期讨论中,卡米利和谢泼德指出,项目反应理论十分适于识别项目功能差异:

> 在研究项目功能差异时,项目反应理论与传统测量理论相比,具有显著优势。首先,项目反应理论对于项目参数(如难度和区分度)的估计较少受到样本特征的影响。其次,项目的统计属性可以被更精确地描述。因

此，当一个项目在两组间具有功能性区别时，这种区别也可以被更精确地描述。最后，项目的统计属性在项目反应理论框架下可以被更好地用图表表现出来，这可以帮助加深对具有功能差异的项目的理解（Camilli & Shepard，1994:47）。

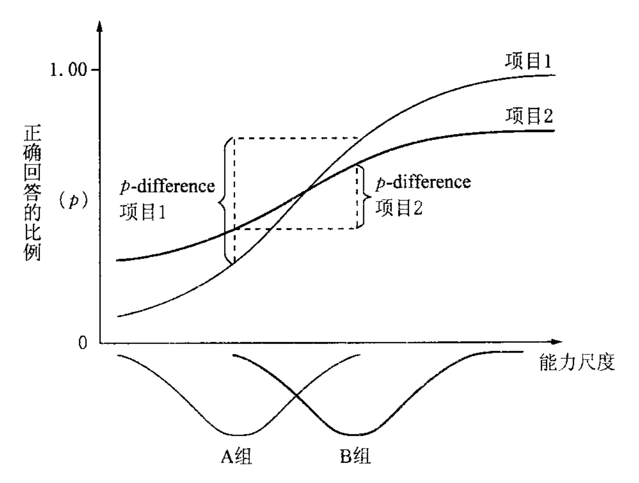


图9.1　两个具有不同项目反应函数的组的比较

图9.1呈现了以上论述。此图包含两个项目的项目特征曲线。在此，我们只考虑项目特征的一个方面，即难度。在能力轴（θ）上，我们给出了两个假设组别（对照组和焦点组）的分布。由组间平均值的差别可以看出，假定两个项目都存在功能差异，回答正确的概率存在区别，并且项目1的区别较项目2更加明显。然而，仅是项目难度上的差异并不一定意味着这差异是显著的，要下这个结论，必须满足其他条件。因此，在项目功能差异研究中使用项目反应理论方法，其目的在于，衡量一个项目是否能够相似地检测不同组

别的被测者的隐性能力。

在这一框架下,在一个给定项目上回答正确的可能性,即是被测者能力(θ)与一个或多个项目特征与项目参数的函数(若想更详细地了解依托项目反应理论的项目参数计算方法,可参见 Baker, 2001)。多年来,研究者发展出了一系列项目反应理论模型,以适用于不同的操作情况(在此测量语境下,模型指的是对项目特征和被测能力在给定项目回答正确的可能性的关系的数学描述)。通常来看,这些模型以包含多少项目特征区分。三个最常见的项目反应理论模型分别为一参数、二参数和三参数项目反应理论模型(简称为 1PL、2PL 和 3PL)。在下文中,我们将描述这些模型及其包含的假设。当然,在开始描述模型前,我们首先需要讨论其项目特征曲线和项目反应函数的作用。

第2节 | 项目反应曲线

在实际操作中，由项目反应理论模型计算出的函数通常表现为一个项目在被测者能力和回答正确概率两轴上的曲线。由于它来源于被测者可观测的反应，因此这一曲线又被称为项目特征曲线或项目反应函数。尽管存在许多项目反应理论模型，但在本部分，我们只讨论由1PL、2PL和3PL模型生成的项目特征曲线。如先前图9.1以及此处的图9.2所示，阐述项目特征曲线最简便的方法是使用图表。X轴为“能力”，Y轴为“回答正确的概率”，$p(\theta)$。图像为一条平滑的项目反应函数曲线，严格意义上说，是一条累积曲线。

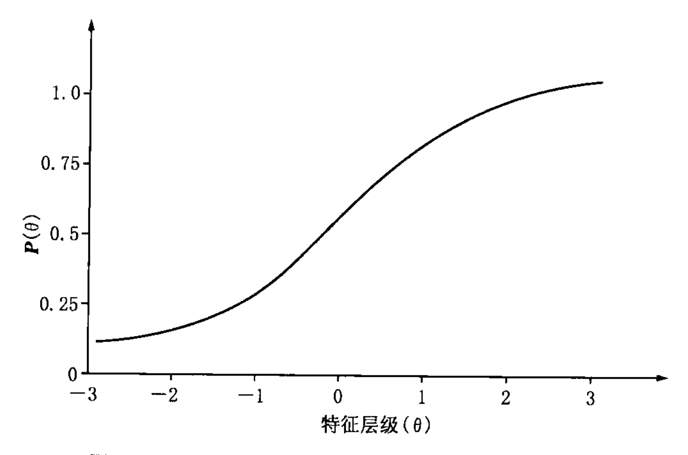


图9.2 能力作为特征层级的函数的描述性项目累计曲线

在项目反应理论中，能力尺度被归一化，即中点为0，两边对称。如其他正态频率分布一样，X轴可以包含正无穷到负无穷。然而，这样的区间在实际操作中并不常见，因而，通常正负三标准差的区间就足以描述总体的全部能力区间。相应地，Y轴表示的$p(\theta)$值为0到1之间。这一量表表示的是在区间(0，1)范围内，从完全为0到完全确定的概率估计。因此，这一曲线描述的是一个给定项目的特征和被测者能力间的函数关系。尽管通常情况下这条累积曲线是S形，但在有些情况下，曲线也可能是其他形状。然而，在所有情况下，累积曲线单向增大，即斜率总是不断增大的。在经典的S形中，曲线从低处开始，逐渐上升，并在接近上限时变缓。然而，曲线从不真正到达上下限。曲线在上下限时变缓被称为渐近线。在项目特征曲线中，有两条渐近线，分别在无限接近上限和下限时。渐近线即为函数曲线和上下线相切时。因此，函数曲线无限接近，却不真正到达上下限。上限通常不在图表上画出来，因为它的定值为1，很容易想象。

现在我们来讨论图9.3中的项目特征曲线。图9.3同时描述了三个项目的项目反应函数，因此我们可以看出其中的差别。三个函数的下渐近线是不同的，项目1约为0.1，项目2略微大于0.25，项目3接近0.5。下渐近线代表着“回答正确的概率”的起始点，因而对于低能力的被测者而言意义重大。在起始点下无测量值。

对于不同起始点的分析为：项目1对于低能力的测试者更难(只有10%的可能性回答正确)，相比而言，对于相同能力的测试者而言，有50%的可能性正确回答项目3。对于低

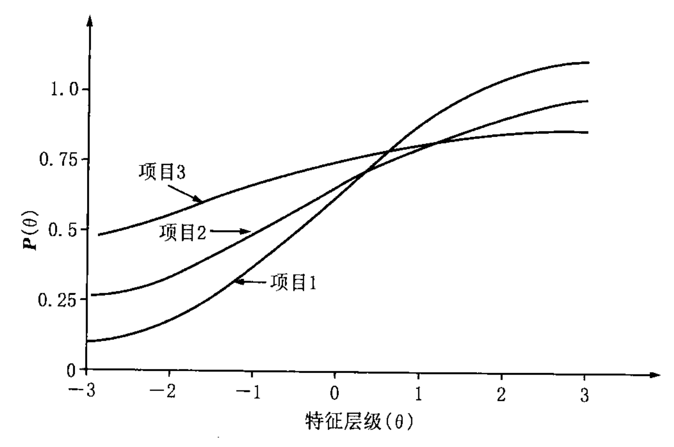


图 9.3 三个项目反应函数的累计曲线

能力的测试者而言,这些项目的难度是不同的。由于此时被测者获得正确答案的概率很低,因而起始点通常也被称为"猜测值"(guessing value)。然而,当我们观测另一条渐近线时可以发现,对于能力高的被测者而言(θ 值大于 2 的被测者),对于渐近线的阐释是类似的。项目的难易度显而易见。

对于项目反应理论而言,项目特征曲线中尤其重要的一个方面是其拐点,即上下渐近线的中点,此时斜率最大。在此,当 c 参数(后文将详细解释)为 0 时,回答正确的概率为 0.5(当 c 参数不为 0 时,拐点位于概率为 $1+c/2$ 处)。因此,对于一个容易的项目,在 θ 值低端,回答正确的概率极可能发生改变(高于 0.5);而对于难的项目,改变可能发生在 θ 值较高处。当被测者的能力值于 X 轴上的位置恰好也处于拐点时,此时项目的难度完美地契合了被测者的能力。图 9.4 是图 9.2 中的三条项目特征曲线,并标出了拐点的位置。项目 1 的拐点位于−0.5 处,而项目 2 和项目 3 的拐点分别位

于 0 和 1 处。

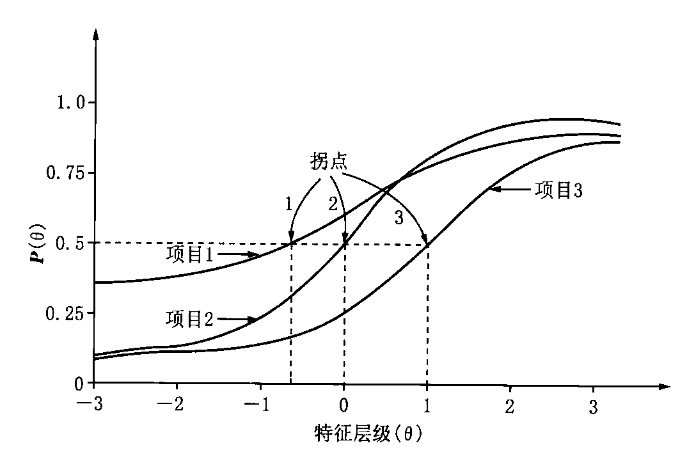


图 9.4 在特征轴上标注出拐点的三个项目的累计曲线

在项目反应理论和项目功能差异中研究这些分界点十分必要。首先,它显示出项目区分度最高对应的能力值。当项目区分度小于 0.5 时,并不能为衡量能力提供太多信息,因为被测者可能觉得项目太过容易或者过于困难,无法进行有效测量。据此,我们可以衡量将一个项目给予被测者是否合适。其次,分界点也能为项目的难度提供信息。下面,我们将讨论三个常用的项目反应理论模型。

第3节 | 单参数模型

单参数模型又称Rasch模型，根据丹麦数学家乔治·拉什（Georg Rasch）命名。在单参数模型中，只有估计项目难度一个参数（在这个模型中，项目的区分度被视为一个定值常数）。单参数模型的核心心理测量学考量的是一个给定项目对于测试者的难度。当被测者根据一定方式回答问题时，其隐性能力就显现出来了。在这一模型中，回答正确即意味着能力更高，回答错误则意味着能力较低。尽管单参数模型具有理论意义，然而在实际操作中却比较困难。因为它对测试项目做出了太多假设，假设项目反应仅仅是项目难度的函数，而与其他任何方面无关。公式9.1为单参数模型中被测者获得正确答案的概率：

$$\frac{\theta_a^*}{b_i^*} \qquad [9.1]$$

*号代表能力并非独立于项目难度。此外，一个事件出现的概率被定义为$P/(1-P)$。在测试中，获得正确答案的几率为获得正确答案的概率与获得错误答案的概率的比例。因此，公式9.2表示了用获得正确答案的概率比值比反映的能力概率。

$$\frac{\theta_a^*}{b_i^*} = \frac{P_i(\theta_a)}{1-P_i(\theta_a)} \qquad [9.2]$$

公式 9.2 可以被简化为公式 9.3：

$$P_i(\theta_a) = \frac{\theta_a^*}{\theta_a^* + b_i^*} \qquad [9.3]$$

单参数模型的两个核心假设为:项目的区分度始终是一致的,猜测参数(guessing, pseudochance parameter)为 0。因此,根据公式 9.4 和公式 9.5 设置被测者能力和项目难度后,公式 9.6 即为单参数模型:

$$\theta_a^* = e^{D_{\bar{a}}\theta_a} \qquad [9.4]$$

$$b_i^* = e^{D_{\bar{a}}b_i} \qquad [9.5]$$

$$P_i(\theta_a) = \frac{e^{D_{\bar{a}}\theta_a}}{e^{D_{\bar{a}}\theta_a} + e^{D_{\bar{a}}b_i}} \qquad [9.6]$$

此外,单参数模型假设获得正确答案个数相同的被测者具有等同的能力,不考虑它们获得正确答案的项目是否一样。此外,公式 9.6 等同于公式 9.3。因此,单参数模型可以被视为传统项目反应理论模型中的一员。

第4节 | 双参数模型

在20世纪50年代及之后，伯恩鲍姆(1958、1968)提出了一个非线性的回归模型，用以描述双参数。这两个参数分别为项目区分度与项目难度。观察参数图像可以帮助我们了解这两个项目的特点。图9.5即为项目1和项目2包括了这两个参数的项目特征曲线。

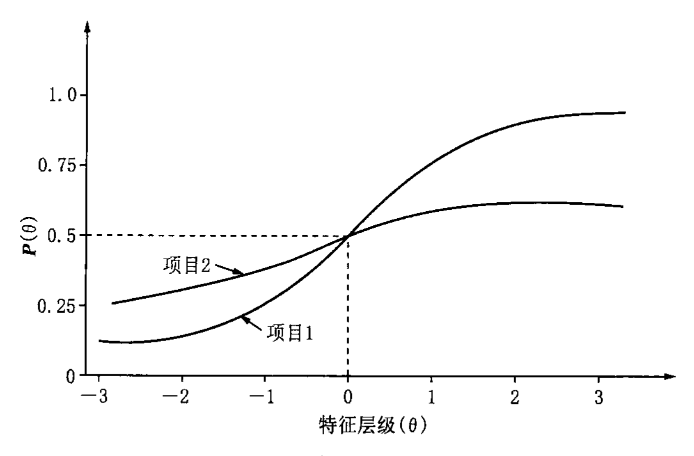


图9.5　具有相同拐点不同斜率的项目特征曲线

两条项目特征曲线并不完全相同，尽管拐点都在 $\theta=0$、$P=0.5$ 处，两条曲线具有不同斜率。从图像上可以看出，随着 θ 增加，获得正确答案的概率也增加。两条曲线斜率的区

别反映的是对于不同能力的被测者,不同项目的区分度不同。对于项目 1 而言,对于低能力的被测者,区分度是渐进的,当 $\theta=-1$ 时,项目的区分度显著提高,直到 $\theta=1$ 时,对于高能力的被测者而言,项目的区分度再一次减弱。将项目 1 与项目 2 对比可以发现,尽管项目 1 和项目 2 曲线的斜率走势相似,但项目 2 在 θ 轴的所有位置,区分度都较为不明显。这一项目特点被称为区分度参数,即 a 参数。当一个项目的斜率在某一个 θ 值格外倾斜时,相应的项目特征曲线将具有 Guttmann 量表的形状(Guttmann, 1950)。对于量表和测试研发而言,通过项目特征曲线研究测试项目的区分度十分必要。

图 9.6 为另外两个具有不同特征的项目的项目特征曲线。在此我们将它们表示为项目 3 和项目 4。由图可见,在中部,项目 3 和项目 4 的斜率近乎相同,然而两条曲线的拐点不同。也就是说,曲线从左到右的变化区分开了项目 3 和项目 4。因而对于相同能力的测试者而言,尽管项目 3 和项

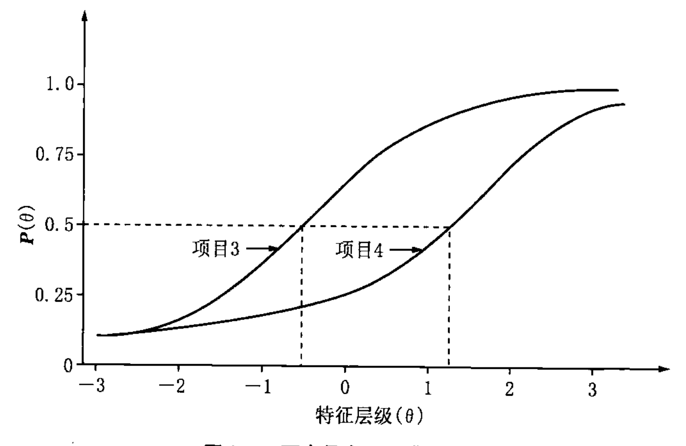


图 9.6　两个具有不同难度项目

目 4 的区分度相似，但项目 4 更加困难，因为它在 θ 轴上的值更大。项目特征曲线在 θ 轴同一位置上的差别意味着难度上的区别，因而，这一参数也被称为难度参数，即 b 参数。

有些时候，在项目反应理论中，项目的难度也被称为位置参数。因为它位于概率函数等于 0.5[$P_i(\theta) = 0.5$]的 θ 值处。

第5节 | 三参数模型

和单参数、双参数模型相比，三参数模型使用并不广泛。然而，对于全面理解项目特征曲线而言，了解三参数模型还是十分必要的。在三参数模型中，包含的参数为项目层面的参数——项目区分度(a)和项目难度(b)，以及猜测容易度(susceptibility to guessing)(c)。第三个项目特征是对双参数模型的一个补充，试图模拟低能力被测者的一个项目选择策略。这一参数，c 参数被称为假机会(pseudochance)水平，但通常称为“猜测参数”。公式 9.7 是三参数模型的完整形式。公式中除了所增加的 c 参数外，其余所有部分的定义同公式 9.6。这一公式描述的是回答正确的概率作为能力(θ)的函数的 S 形曲线。曲线的具体形状受 a、b、c 参数影响。和前面所讨论的曲线一样，这样的曲线通常被称为项目特征曲线或项目反应函数。常数 D 的定值为 1.7，其功能是使得三参数模型的曲线接近普通累积函数的曲线(Osterlind, 2006)。

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}\ (i = 1, 2, \cdots, n) \qquad [9.7]$$

在三参数项目反应理论模型的项目特征曲线中，c 参数为下渐近线。严格来说，c 参数是在 θ 值最小时，获得正确答

案的概率。从图像上说，c 参数在图形左侧，是累积函数的起始点。三参数回归模型是使用最为广泛的项目反应理论模型，特别是对于二元测试项目而言。公式 9.7 为三参数模型的数学式。

在三参数模型中，a 参数是项目的区分度，b 参数是测试项目的难度，c 参数是项目的“假机会”参数，即一个低能力测试者获得正确答案的几率。图 9.7 是一个最典型的完整的三参数模型的图像形式。

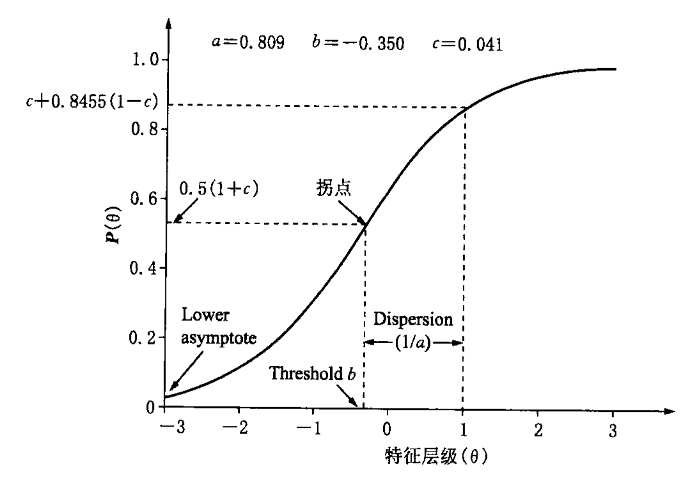


图 9.7 一个项目三参数模型中的尺度调整

第6节 | 依托于项目反应理论的项目功能差异方法

依托于项目反应理论的项目功能差异方法可以有效地检测组间项目特征的区别。正如克劳泽和梅热(1998)等研究者(如 Camilli & Shephard, 1994; Penfield & Camilli, 2007; Zumbo, 1999)提出的,项目反应理论所具有的缩放性质可以帮助我们用简洁明了的方式研究一个特定的测试项目。尽管存在一系列识别项目特征组间差别的依托于项目反应理论的方法,但没有任何一个方法是完美的。同时需要注意的是,在项目反应理论框架下,有两种对于项目功能差异的理解方式:项目特征曲线中的区别和项目参数中的区别。彭菲尔德和卡米利于2007年的研究提醒我们,这两种理解都帮助我们发展了一系列项目功能差异的研究方法。总的来说,几乎所有统计方法都有近乎相同的原假设,即项目参数在焦点组和对照组间是恒定的(Clauser & Mazor, 1998)。

第7节 | 项目参数中的区别

一个更直接的检验项目功能差异的方法是比较焦点组和对照组的项目参数,尤其是 b 参数。根据洛德(1980)的研究,公式 9.8 是最简单的测试"焦点组和对照组 b 参数相同"这一原假设的方法。

$$d=\frac{(\hat{b}_R-\hat{b}_F)}{SE(\hat{b}_R-\hat{b}_F)} \qquad [9.8]$$

其中 $SE(\hat{b}_R-\hat{b}_F)$ 是焦点组和对照组 b 参数区别的标准误。见公式 9.9:

$$SE(\hat{b}_R-\hat{b}_F)=\sqrt{[SE(\hat{b}_R)]^2+[SE(\hat{b}_F)]^2} \qquad [9.9]$$

这一统计量 d 的分布近乎正态分布,因此,是对于原假设 $b_R=b_F$ 的测试。

的确,如果单参数模型适用于我们的数据,那么统计量 d 提供了一个最简单直接的对原假设(即不存在项目功能差异)的检验。然而,如果数据更适合双参数或三参数模型,那么仅仅检验 b 参数中的区别就有失偏颇了。在这种情况下,根据洛德(1980)的研究,我们建议使用同时检验 b 参数和 a 参数的卡方检验。在这种情况下,参数中的区别见公

式 9.10。

$$\hat{v}' = (\hat{a}_{\mathrm{R}} - \hat{a}_{\mathrm{F}}, \hat{b}_{\mathrm{R}} - \hat{b}_{\mathrm{F}}) \quad [9.10]$$

相应的统计量可以被计算为:

$$\chi_{\mathrm{L}}^{2} = \hat{v}' S^{-1} \hat{v} \quad [9.11]$$

在公式 9.11 中,S 代表 a 参数和 b 参数组间差别的方差—协方差矩阵。因此,假设不存在项目功能差异,所获得的 χ_{L}^{2} 是一个自由度为 2 的卡方分布。有兴趣的读者可以进一步研读洛德(1980)与卡米利和彭菲尔德(1994)的研究,以深入了解通过检验项目参数检验项目功能差异的方法。

第 8 节 | 似然比检验

似然比检验也被用于依托于项目反应理论的项目功能差异研究。这一检验比较一个参数在焦点组与对照组间相等的可能性和不等的可能性。根据彭菲尔德和卡米利(2007)近期的描述,将项目参数限制为在焦点组和对照组间相等的似然性定义为 $L(C)$,而将项目参数可以在组间不同的似然性定义为 $L(A)$,公式 9.12 即为似然性检验公式:

$$G^2 = 2\ln\left[\frac{L(A)}{L(C)}\right] \qquad [9.12]$$

正如其他研究者提到的(见 Thissen、Steinberg & Wainer, 1993),此统计量分布近似于卡方分布。感兴趣的读者可进一步阅读这些资料。

从实践操作角度讲,对项目功能差异感兴趣的研究者通常检验 b 参数在焦点组和对照组间是否相同。在使用应用广泛的、估计项目反应理论参数的软件时(例如 BILOG-MG),a 参数虽被估计,但被限制为在组间是相等的。需要再次强调的是,这里需要考虑项目在 θ 轴上的位置。相应地,当使用 BILOG-MG 研究项目功能差异时,生成的是单参数模型。为了阐释清楚,图 9.8 给出了使用 BILOG-MG 研究大学 BASE 测试项目功能差异的语法示例。关于语法的进一步信

息，以及有关FORTRAN命令的信息，参见BILOG-MG使用说明(du Toit, 2003)。

```
College BASE data-Form LP
>COMMENT
Sample run with CB data, form LP for DIF. Two groups are used: Male & Female
The needed files (beyond this *.blm file are labeled:
          the data file                    CB2data.dat
          the answer key                    CB2KEY.Key
          the omit file (coded as 9)        CB2omit.omt
Scoring phase is not invoked as it is invalid for DIF analysis; but PLOT command is used.
After the data is run, use Run-->Plot to activate the graphics module.
>GLOBAL DFName = 'C:\Program Files\bilogmg\Examples\CB2data.dat',
        NPArm = 1;
>LENGTH NITems = (41);
>INPUT NTOtal = 41, NALt = 5, NIDchar = 8, NGRoup = 2,
       KFName = 'C:\Program Files\bilogmg\Examples\CB2key.KEY',
       OFName = 'C:\Program Files\bilogmg\Examples\CB2omit.OMT',
       DIF;
>ITEMS INAmes = (ITEM001(1)ITEM041);
>TEST1 TNAme = 'CB_LP', INUmber = (1(1)41);
>GROUP1 GNAme = 'MALES', LENgth = 41, INUmbers = (1(1)41);
>GROUP2 GNAme = 'FEMALES', LENgth = 41, INUmbers = (1(1)41);
(8A1, 3X, I1, 33X, 42A1)
>CALIE PLot = 1.0000;
```

图9.8　项目功能差异的BILOG-MG命令

总的来说，这些项目参数由不同的项目特征曲线表示，焦点组和对照组各一条。图9.9到图9.14是同一个BASE测试中同一个项目的项目功能差异相关统计量。图9.9是一个项目焦点组和对照组(女性在下方，男性在上方)的项目特征曲线及相应项目信息。

特别需要重视的是，b值的组间差别很大。项目特征曲线中的显著差别可以被视为存在项目功能差异的证据(Lord, 1980)。感兴趣的读者可以阅读西森等学者(Thissen et al., 1993)的著作以了解计算这些值的数学方法。西森在许多其他计算机软件中(MULTILOG, LISCOMP, SPSS

LOGLINEAR，LOGIMO，BIMAN)也描述了它的运行。

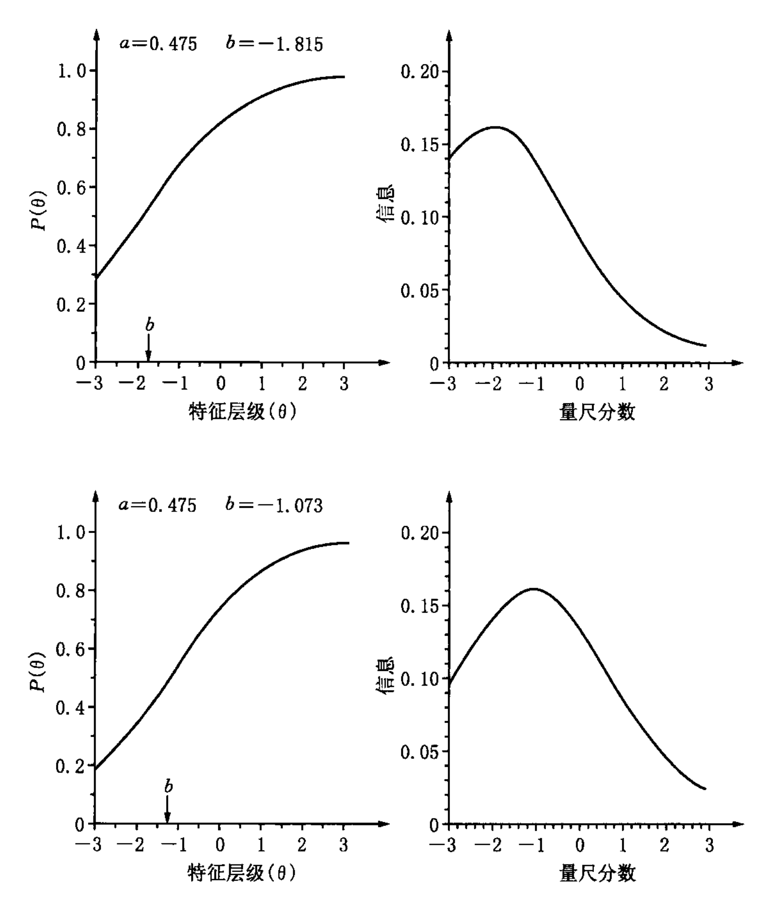


图 9.9 对照组(上图)和焦点组(下图)的项目特征曲线

其他依托于项目反应理论的方法也计算组间统计量。边缘极大似然估计法可用 BILOG-MG 进行，具体见图 9.10 和图 9.11。图 9.10 是传统的项目系数(各组的和全样本的)。然后是依据项目反应理论的参数估计。

ITEM STATISTICS FOR GROUP: 1 MALES

ITEM	NAME	# TRIED	# RIGHT	PCT	LOGIT/1.7	ITEM * TEST CORRELATION PEARSON	BISERIAL
1	ITEM01	953.0	748.0	0.785	-0.76	0.181	0.254
2	ITEM02	953.0	792.0	0.831	-0.94	0.187	0.278
3	ITEM03	953.0	818.0	0.858	-1.06	0.152	0.236

ITEM STATISTICS FOR GROUP: 2 FEMALES

ITEM	NAME	# TRIED	# RIGHT	PCT	LOGIT/1.7	ITEM * TEST CORRELATION PEARSON	BISERIAL
1	ITEM01	3 933.0	2 690.0	0.684	-0.45	0.236	0.309
2	ITEM02	3 933.0	3 272.0	0.832	-0.94	0.204	0.304
3	ITEM03	3 933.0	3 511.0	0.893	-1.25	0.190	0.319

ITEM STATISTICS FOR MULTIPLE GROUPS CB_LP

ITEM	NAME	# TRIED	# RIGHT	PCT	LOGIT/1.7	ITEM * TEST CORRELATION PEARSON	BISERIAL
1	ITEM01	4 886.0	3 438.0	0.704	-0.51	0.228	0.301
2	ITEM02	4 886.0	4 064.0	0.832	-0.94	0.200	0.298
3	ITEM03	4 886.0	4 329.0	0.886	-1.21	0.180	0.296

图 9.10 对照组和焦点组的经典项目统计量

GROUP 1 MALES; ITEM PARAMETERS AFTER CYCLE 3

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF
ITEM01	0.862 0.048*	0.475 0.004*	-1.815 0.101*	0.429 0.004*	0.000 0.000*	5.4 (0.798 3)	9.0
ITEM02	1.054 0.052*	0.475 0.004*	-2.220 0.110*	0.429 0.004*	0.000 0.000*	2.5 (0.981 0)	9.0
ITEM03	1.186 0.056*	0.475 0.004*	-2.498 0.118*	0.429 0.004*	0.000 0.000*	4.4 (0.880 3)	9.0

GROUP 2 FEMALES; ITEM PARAMETERS AFTER CYCLE 3

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF
ITEM01	0.509 0.021*	0.475 0.004*	-1.073 0.045*	0.429 0.004*	0.000 0.000*	4.7 (0.857 8)	9.0
ITEM02	1.040 0.026*	0.475 0.004*	-2.191 0.055*	0.429 0.004*	0.000 0.000*	1.0 (0.999 4)	9.0
ITEM03	1.364 0.031*	0.475 0.004*	-2.873 0.066*	0.429 0.004*	0.000 0.000*	6.7 (0.671 9)	9.0

图 9.11 对照组和焦点组依托于项目反应理论的项目统计量

为进行有意义的比较，对照组根据项目反应理论的临界值进行了线性变换，之后焦点组在度量上相同。这一点十分重要，具体见图 9.12。

PARAMETER		MEAN	STN DEV	
GROUP:	1	NUMBER OF ITEMS:		41
THRESHOLD		−1.002	0.928	
GROUP:	2	NUMBER OF ITEMS:		41
THRESHOLD		−0.877	0.944	

THRESHOLD MEANS

GROUP	ADJUSTMENT
1	0.000
2	0.125

图 9.12　项目反应理论参数估计的调整

然后，计算调整过的焦点组和对照组的临界值（见图 9.13）。最后，图 9.14 是焦点组与对照组的区别。区别的值被看做相对量，区别越大，则意味着存在项目功能差异。

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
ADJUSTED THRESHOLD VALUES

ITEM	GROUP 1	GROUP 2	ITEM	GROUP 1	GROUP 2
ITEM01	−1.815	−1.198	ITEM22	−2.112	−2.005
	0.101*	0.045*		0.113*	0.052*
ITEM02	−2.220	−2.315	ITEM23	−0.376	−0.312
	0.110*	0.055*		0.088*	0.043*
ITEM03	−2.498	−2.998	ITEM24	−0.695	−0.731
	0.118*	0.066*		0.088*	0.043*

图 9.13　调整后的阈值

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING:
GROUP THRESHOLD DIFFERENCES

ITEM	GROUP 2-1	ITEM	GROUP 2-1	ITEM	GROUP 2-1
ITEM01	0.618	ITEM15	-0.405	ITEM29	0.079
	0.110*		0.095*		0.101*
ITEM02	-0.096	ITEM16	0.007	ITEM30	-0.059
	0.135*		0.096*		0.113*
ITEM03	-0.500	ITEM17	-0.182	ITEM31	0.354
	0.135*		0.095*		0.137*

图 9.14 阈值调整后的组间区别

一个传统的用以检验这些数值的方法是学生 t 检验。这里，被除数是给定项目 b 参数的组间差别。除数是两个标准误之和的平方根。见公式 9.13(S. E. Embretson，私人通信，2003 年 4 月)。

$$t_{\mathrm{DIF}} = \frac{b_{\mathrm{F}} - b_{\mathrm{R}}}{\sqrt{SE_{\mathrm{F}} + SE_{\mathrm{R}}}} \qquad [9.13]$$

b_{F} 是焦点组的难度参数，b_{R} 是对照组的难度参数。

另一个计算项目功能差异的方法是将焦点组调整过的 b 值与对照组的 b 值相减。如果其差值大于两到三个标准误，则可认为这一差异是显著的。正如我们前面所提到的，将两个标准误的组间区别作为显著标准可能导致过度判断项目功能差异。因而，有些研究者使用更为保守的标准(三个标准误)。但无论如何，这一方法都是可行的，因为根据 BILOG-MG，区别分数是标准化的 z 分数。

第9节 | 区域测量

对影响大小或统计显著性的估计是根据比较项目参数确定的。可以通过比较焦点组与对照组项目特征曲线的面积差别，量化两组间在难度参数和区分度参数上的区别(Raju, 1988; Rudner Gagne & Gagne, 2001)。在有些情况下，甚至仅仅通过看图就可以做出一个合理的判断，然而更好的是一个使用一定统计方法的、更准确的计算。一个常用的方法，即有向面积影响大小指数，最早由鲁德(Rudner)和加涅(Gagne)于2001年提出，并由拉古(Raju)、文·德·林登(Van der Linden)和弗利拉(Fleer)进一步发展。具体见公式9.14。

$$SA = \int_{-\infty}^{\infty} [P(Y=1 \mid \theta, G=R) - P(Y=1 \mid \theta, G=F)] d\theta \qquad [9.14]$$

假设组间 c 参数保持一致，有向面积可以被看做焦点组和对照组的项目参数函数。见公式9.15:

$$SA = (1-c)(b_F - b_R) \qquad [9.15]$$

在假设 c 参数一致后，有向面积独立于 a 参数，即便 a 参数并非在组间不变(Penfield & Camilli, 2007; Raju, 1988)。项目功能差异是组间项目难度参数差异的函数。此外，组间 b

参数差异越大,导致组间给出正确答案的概率的差异也越大,即项目功能差异影响越大。因而,对于单参数模型而言,有向面积等于两组的难度参数差异。

然而尽管如此,研究者必须谨慎使用这一方法。正如彭菲尔德和卡米利指出的,有向面积在 a 参数并非组间无差异的情况下,可能有失偏颇。特别是项目特征曲线在 θ 于 -3 到 3 区间内有交叉的情况下,更是如此。在这种情况下,项目特征曲线的组间差别在一定范围内是正向的,在其他范围内是负向的。这时,即便组间差别仍是显著的,但有向面积却可能较小。

为了解决这一问题,拉古与同事(1988)提出考虑组间无向面积的区别,见公式 9.16。

$$\mathrm{UA} = \int_{-\infty}^{\infty} [P(Y=1 \mid \theta, G=R) - P(Y=1 \mid \theta, G=F)] d\theta \tag{9.16}$$

当 c 参数在组间恒定时,无向面积等于:

$$\mathrm{UA} = (1-C)\left| \frac{2(a_{\mathrm{F}} - a_{\mathrm{R}})}{Da_{\mathrm{F}}a_{\mathrm{R}}} \ln\left[1 + \exp\left(\frac{Da_{\mathrm{F}}a_{\mathrm{R}}(b_{\mathrm{F}} - b_{\mathrm{R}})}{a_{\mathrm{F}} - a_{\mathrm{R}}}\right)\right] - (b_{\mathrm{F}} - b_{\mathrm{R}}) \right| \tag{9.17}$$

当 a 参数也在组间恒定时,公式 9.16 可进一步被简化为:

$$\mathrm{UA} = (1-C) \mid b_F - b_{\mathrm{R}} \mid \tag{9.18}$$

有向面积和无向面积从理论上说,都可以衡量项目功能差异的大小。然而它们却不能有效地检验不存在项目功能

差异这一原假设。面积测量的局限性在于没有考虑到被测者在 θ 轴上的分布，因而对于特定种族或阶层的被测者，可能有失偏颇。此外，有向面积和无向面积需要分开估计各组的项目参数，因而对于数据有特定的要求——因为尤其是对于焦点组而言，样本大小可能不足以稳定计算参数值。尽管如此，在此我们仍然根据彭菲尔德和卡米利的研究对它做出介绍：

> 有向面积和无向面积描述了项目功能偏差大小和组间项目参数差异间的重要理论关系。这一关系常被用来理解项目功能差异分析的结果，并应对研究中存在不同程度的项目功能差异的情况(Penfield & Camilli, 2007:131)。

第10节 多分类项目中检验项目功能差异的项目反应理论方法

和二分类项目一样，关于检验多分类项目的功能差异，同样存在大量文献(例如 Penfield & Lam, 2000; Potenza & Dorans, 1995)。许多是二分类项目检验方法的延伸，尽管从理论上说，多分类项目中的功能差异并不是同样简明直接的。但对于多元测试项目，项目反应和分组变量 G 的关系可能是任意一个反应的组间区别的函数。而对于二分类项目而言，项目反应理论框架对于估计一个反应作为 θ(被测者能力)的函数十分有用。目前提出的许多多分类项目检验模型，都是拓广的分部评分模型(GPCM)或等级反应模型(GRM)的变体。GPCM 模型(Muraki, 1992)在操作中广泛使用。GPCM 估计获得反应 j 的可能性，根据双参数模型，有：

$$n_j(\theta) = \frac{\exp[Da(\theta - b_j)}{1 - \exp[Da(\theta - b_j)} \qquad [9.19]$$

根据彭菲尔德和卡米利(2007)的注释方法，简言之，对于一个有 r 种反应的测试项目，我们将一个特定的反应定义为 j，$j = 0, 1, \cdots, J$。$J = r - 1$。有 J 个相邻的反应概率函数。因而，b_j 参数估计的是从 $j-1$ 到 j 的难度。

相似地，鲛岛研一郎(Samejima，1997)的GRM模型将反应选择$Y \geqslant j > 0$的总概率阐释如下：

$$y_j(\theta) = \frac{\exp[Da(\theta - b_j)}{1 - \exp[Da(\theta - b_j)} \qquad [9.20]$$

和村木(Muraki，1992)的GPCM模型一样，鲛岛研一郎的GRM模型假定J为总概率函数。随着b_j参数增加，获得一个j反应所需要的θ值也增加。正如我们之前提到的，似然比检验同样可以通过MUTILOG7进行(Thissen、Chen & Bock，2003)。然而，博尔特(Bolt，2005)在研究Monte Carlo法时指出，似然比检验对于GRM的局限在于可能造成一类错误的增加，特别是在数据和GRM不完全契合时。

第10章

logistic 回归

项目功能差异研究中的 logistic 回归，涉及一个使用最大似然估计法的概率函数(见 Hosmer & Lemeshow, 1989)。在项目功能差异方法中，因变量是定类变量，意为对于一个特定项目给出特定答案(对于二分类项目来说，即回答正确或错误;对于多分类项目来说，答案是序数)的可能性。项目反应控制组别(对照组和焦点组分别被定义为两个哑变量)。其他的因变量为被测者的能力变量，以及一个交互项。正如之前讨论 M-H 方法时提到的，P 和 Q(即 $1-P$)分别对应回答正确和错误的被测者的比例。这意味着因变量的方差等于 $P(1-P)$(即 PQ)。同时，由于每一层 Y 值的总体均值和 X 各值并非一条直线，两者的关系并非线性。此外，误差并不是正态分布的，因为关于齐方差性的假设无法满足。因而，无法通过普通最小二乘法(Ordinary Least Squares, OLS)合理估计总体参数。因而，我们需要经过对数转换，用最大似然估计法进行参数估计。

以上即 logistic 回归的主要特点，它们在项目功能差异研究中十分有用。它的一个优势在于，将一个项目的特定反应(回答正确或错误)表达为一个概率，与教育和心理测量中对结果的解读的一贯方法是一致的。我们需要谨记，尽管有

着一系列精确的统计方法，但精神评估却并非一项精准的科学。因而“概率”是一个更加合理的说法。此外，在项目功能差异中使用 logistic 回归也有操作上的益处。例如，关于数据的正态分布假设可以放松，因而在一系列现实生活中的小样本情况里，logistic 回归都有用武之地。

在项目功能差异研究中使用 logistic 回归的另一大优势在于，研究者可以使用这种方法研究统一的功能差异和不统一的功能差异。此外，这一方法也不局限于二分类项目。多分类项目也可以用这一方法研究，尽管常常局限于三组的情况（见 French & Miller, 1996）。尽管如此，这一方法广泛适用于李克特(Likert)和类李克特量表项目和短文评分的各种情况。

然而，需要特别注意的是，logistic 回归关注的是评分模型(例如二元或多类)，而非项目的形式本身。换言之，在二元回归中，项目的形式——无论是多项选择、正误判断或其他二元评分形式——都是无所谓的。对于多分类项目而言，从操作角度说，定类的因变量支持两组或三组的情况，尽管回归理论中理论上允许多于三组的情况出现。因而，当用 logistic 回归分析李克特或类李克特量表的项目反应的项目功能差异时，研究者可能需要聚合一些反应，以保证分析中只包括两到三个类别。

最后，和上文提到的心理学解读一致，运用 logistic 回归分析项目功能差异同样遵循心理测量中的一个基本假设，即，被测者在项目上的应答反映的是被测者的隐性能力渐变。即便用以测评的项目是二元的，这种隐性能力的渐变仍然存在。M-H 方法、项目反应理论和项目反应理论模型等分析，都依赖于这一基本假设。

第 1 节 | logistic 回归的项目功能差异表达

下面,我们来解释一下运用 logistic 回归进行项目功能差异的基本方法。回归公式如 10.1 所示:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \qquad [10.1]$$

显然,Y 为因变量。在公式右边,β 为各自变量的系数。截点为 β_0,其含义是,当 X_1 和 X_2 为 0 时,出现一个反应类别的可能性。第一个自变量 X_1 是配对依据,通常是测试总分。第二个自变量 X_2 是组别,这是一个哑变量。交互项用 $X_1 X_2$ 表示。

第一个自变量上的概率差异意味着统一的项目功能差异。不统一的项目功能差异表现为一个不为 0 的交互项。这里的道理与在分析中将能力视为分散的无序类别一样。

当公式 10.1 中的因变量被表达为比值比的自然对数时,即为公式 10.2。

$$P(u = 1 \mid \theta) = \frac{e^{(\beta_0 + \beta_1 \theta)}}{[1 + e^{(\beta_0 + \beta_1 \theta)}]} \qquad [10.2]$$

公式左边是根据能力(θ),一个项目回答正确的概率。项目功能差异可以为焦点组和对照组分别用公式表达。见公

式 10.3。

$$P(u_{ij}=1\mid\theta)=\frac{e^{(\beta_{0j}+\beta_{1j}+\theta_{ij})}}{[1+e^{(\beta_{0j}+\beta_{1j}+\theta_{ij})}]}\qquad[10.3]$$

这里，对于个体(i)来说，其概率受其组别(j)影响。这里有两组，即焦点组与对照组。项目功能差异意味着两组间能力相当的测试者获得正确答案的概率不同。也就是说，当回归分析的曲线斜率相等时，即不存在项目功能差异。在公式 10.3 中，这可以表现为 $\beta_{01}=\beta_{02}$ 且 $\beta_{11}=\beta_{12}$。而当 $\beta_{01}\neq\beta_{02}$ 且 $\beta_{11}\neq\beta_{12}$ 时，即存在统一的项目功能差异。而当存在能力和组别的交互，即 $\beta_{01}=\beta_{02}$ 且 $\beta_{11}\neq\beta_{12}$ 时，则意味着存在不统一的项目功能差异。

就分析而言，变量按顺序放入回归公式。首先放入配对依据(即测试总分)，然后是组别变量，最后是交互项。结果是最后的自由度为 2 的卡方分布统计量。自由度 2 是由交互项的自由度(3)和原始的控制变量的自由度(1)的区别计算而来。对于独立性的测试是用以确立配对的项目反应是否与理论分布不同。同样的自由度为 2 的卡方分布用以分析二类分数和定序分数。之所以可以这样，是由于组别变量和控制变量的比较和交互变量与控制变量的比较是平行的。这也可以同时检验统一的项目功能差异和不统一的项目功能差异。

第 2 节 项目功能差异 logistic 回归示例

下面我们通过普通能力测试中的两个项目距离说明项目功能差异中 logistic 回归的使用。这两个项目(项目 4 和项目 12)中,1 为回答正确,0 为回答错误。样本是 300 个被测者,被分为焦点组和对照组。计算使用 SPSS v. 16(SPSS, 2008)。分析的句法参见朱姆伯(1999)。

图 10.1 是一部分分析结果。与前文一样,这里的例子也来自大学 BASE 测试。图 10.1 中包括了交叉统计计算的描述统计。特别需要注意的是,方框 1 中的卡方值和 Nagelkerke 的 R^2。这一检验检验的是统一的项目功能差异。方框 2 中包含了交互项,描述的是不统一的项目功能差异。具体见图 10.2。

Item_4 * group Crosstabulation

Count

		group		
		Reference	Focal	Total
Item_4	0	300	228	528
	1	53	19	72
	Total	353	247	600

Block 1: Method=Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	133.387	1	.000
	Block	133.387	1	.000
	Model	133.387	1	.000

Model Summary

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	306.923[a]	.199	.383

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted		
			Item_4		
			0	1	Percentage Correct
Step 1	Item_4	0	514	14	97.3
		1	50	22	30.6
		Overall Percentage			89.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	Total_score	.247	.027	81.203	1	.000	1.281
	Constant	−21.629	2.255	91.976	1	.000	.000

图 10.1 logistic 回归结果的开始部分

图 10.3 是 SPSS 的语法。这一套命令只需稍做修改，即可用于许多数据。更详细的语法请参见朱姆伯所著《项目功能差异的理论与操作：二分与李克特量表分数中普遍使用的 logistic 回归》一书的配套网站。

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.097	2	.952
	Block	.097	2	.952
	Model	133.484	3	.000

Model Summary

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	306.826[a]	.199	.384

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted		
			Item_4		
			0	1	Percentage Correct
Step 1	Item_4	0	514	14	97.3
		1	50	22	30.6
		Overall Percentage			89.3

a. The cut value is .500.

Variables in the Equation

		B	S. E.	Wald	df	Sig.	Exp(B)
Step 1	Total_score	.248	.049	25.699	1	.000	1.281
	group(1)	−.259	4.848	.003	1	.957	.772
	group(1) by Total_Score	.002	.060	.001	1	.975	1.002
	Constant	−21.594	3.929	30.210	1	.000	.000

图 10.2 logistic 回归结果的结束部分

```
COMMENT Computes Crosstabs for descriptive, and chi-squared test
(evaluate with 2df) for simultaneous evaluation of uniform and non-uni-
form DIF.
CROSSTABS
/TABLES=Item_4 BY group
/FORMAT=AVALUE TABLES
/CELLS=COUNT
/COUNT ROUND CELL.
LOGISTIC REGRESSION VAR=Item_4
/METHOD= ENTER Total_Score /METHOD= ENTER Group Group *
Total_Score
/CONTRAST(Group)=Indicator
/CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
EXECUTE.
```

图10.3 在SPSS17.0里计算项目功能差异logistic回归的命令

第11章

特定的项目功能差异研究方法

前文所描述的项目功能差异研究方法是较为常用的,然而,对于特定的项目研究而言,它们并非仅有的方法。许多情况下,特定的研究方法更为合适,例如多分类评分测试或翻译版的测试。我们在本章中介绍一些这样的方法。

第 1 节 | 多分类评分项目

研究者越来越多地注意多分类评分项目中的项目功能差异(Koretz & Hamilton, 2006)。研究者正从一维度和多维度等视角看待项目功能差异。此外,我们注意到不同种族间测试表现出越来越大的差异,因而这样的探索在第二代测试中格外重要(Lane & Stone, 2006)。在此,我们简要地描述两个项目功能差异中较有发展前途的方向。第一个是一个依托于项目反应理论的方法,成为项目与测试的功能差异(DFIT)。它由拉古及其同事在许多重要的研究中提出(Flowers、Oshima & Raju, 1999; Raju、Oshima & Wolach, 2005; Raju、van der Linden & Fleer, 1995)。另一个较新的方法结合了等效检验的统计方法(Tyron, 2001; Wells、Wollack & Sherlin, 2008),以解决教育测试中项目功能差异方法所遇到的挑战。

此外,在表现测试中(如证书资格测试),多分类评分项目中的异常项目功能十分重要。通常情况下,使用 M-H 方法。然而在分析结果时候必须特别注意其效度问题(Zwick、Donoghue & Grima, 1993)。之前讨论的 logistic 回归方法也可以被延伸到多分类评分项目的研究中来。一个办法是将多分类反应类编重新编为 $J = r - 1$ 的哑变量。这一办法最

先由弗伦奇和米勒(French & Miller, 1996)提出,由彭菲尔德和卡米利(2007)进一步阐述。

类似于阶乘 ANOVA 的结论,检验多分类项目的项目功能差异也涉及同样的问题,即,差异究竟在哪里?对于多分类项目而言,这意味着确认存在项目功能差异的分数层。项目功能差异可能在一层或多层存在。假设一个项目是四点评分,即存在三个分数层,项目功能差异可能在每层或者只在一层存在,或是在不同分数层方向不同。因而,知道项目功能差异究竟存在于何处,不仅对于测试研发者,同时对于测试评估和其他研究都十分必要。

特别是对于多分类评分项目而言,确定项目功能差异存在于哪个分数层,不仅有助于理解组间区别,同时可以帮助对教学大纲进行评估。此外,它甚至能够帮助发掘导致项目功能差异的原因。彭菲尔德和加特莫特(Penfield & Gattamorta, 2009)将这种对组间属性的检验称为“步骤功能差异”(Differential Step Functioning, DSF)。

DSF 方法可以伴随多分类项目或测试在所有传统的缩放方法中使用,不论它们是依托于传统的测试理论或是项目反应理论。在项目反应理论模型中,被测者完成一个特定步骤的概率被计算为 $J=r-1$ 个函数,这里 r 是反应选项的数目,J 是步骤函数的数目。例如,当 $r=4$ 时,$J=3$。此外,根据不同多分类项目的项目反应理论模型的特点,步骤概率函数有多种形式(见 Muraki & Bock, 2003; Samejima, 1997)。然而,最常见的 DSF 使用的是累积模型。

DSF 模型关注每一个步骤焦点组和对照组的差异。当函数被画成散点图时,就可以看出 DSF 的主要特点(见图

11.1)。对于一个四反应项目(三步函数)而言,散点图有 X 轴和 Y 轴。图上有六条曲线(每步两条,一组一条)。在图 11.1 中,曲线十分接近,因而通过观察即可下结论认为不存在 DSF。当需要进行统计检验时,delta 检验可以被用以检验不存在 DSF 的原假设。但是,需要注意的是,第一步和第二步的曲线在 Y 轴上有交叉。我们可以通过观察第一步、第二步和第三步的曲线的接近程度,获得进一步信息。读者可参见彭菲尔德和加特莫特(2009)有关国家教育测试委员会 DSF 模板。

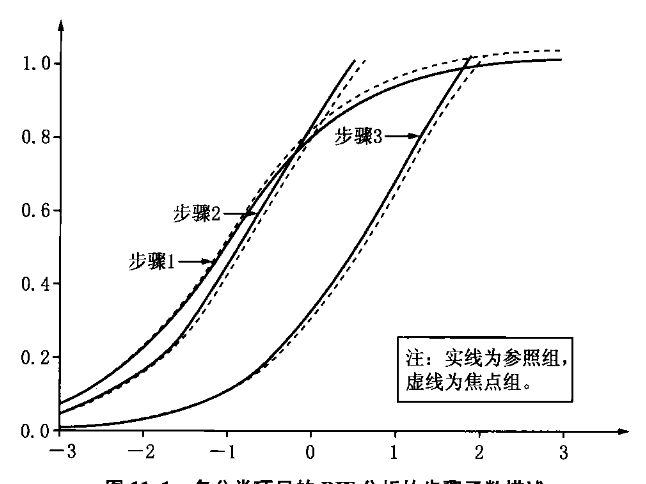


图 11.1　多分类项目的 DIF 分析的步骤函数描述

第2节 | 机考

在许多小规模情况下,使用电脑进行测试十分有用。机考在一些大规模项目中也有一定的成功。通常情况下,在机考中,屏幕上的测试与纸笔进行的测试相同,或近乎相同,并且往往也不包括项目反应理论缩放,以及其他适用于电脑的测试方法(CAT)或"特制"的测试(见下文讨论)。基于电脑的测试可以有许多形式。在电脑上进行测试的优势在于,可以即时评分,并且对测试本身的修正也可以更有效率地进行。而机考的一个核心问题在于测试环境的安全性,因为对测试的管理是远程进行的。

重要的是,无论机考有何优缺点,仍有大量文献探讨机考的问题,认为随着测试媒介的变化,测试也应该从根本上有所不同(见 Mills、Portenza、Fremer & Ward, 2002)。需要注意的是,机考中存在许多心理测量的挑战。若要考虑进行机考,需要了解本领域内最新的研究,以在管理和实践过程中具备必要的知识。若要了解相关文献的元分析,可参见一些学者的著作(Wang、Jiao、Young、Brooks & Olson, 2008)。

尽管有这些问题,但机考中项目功能差异的研究也可以使用我们已经涉及的研究方法。兹维克(Zwick, 2000)提醒研究者,机考形式本身即可能成为误差的来源之一。因此,参与机考的研究者需要在研究焦点组和对照组项目功能差异的过程中考虑机考和纸笔考的区别。

第3节 | 计算机自适应测验(CAT)

计算机自适应测验(CAT)在评价领域内是一个新兴的研究和实践分支。文·德·林登和格拉斯这样描述CAT：

> 和将同样的测试给予所有测试者不同，在CAT中，根据测试者能力，电脑给予被测者不同的测试题。在每一题后，被测者的能力估计都被更新，随后给出的题目可以基于此，最好地估计被测者的能力(Van der Linden & Glas, 2000:vii)。

在这种情况下，被测者被测的测试项目和数目是根据他/她回答之前的项目时所估计的能力。只有对被测者的能力估计较为一致了之后，这一迭代的步骤才停止下来。因为这一原因，CAT中研究项目功能差异更为重要，在这种情况下，有缺陷的项目对于被测者总分的影响也更大(Zwick, 2007)。

有些时候，根据测试项目的多少，比起将同一测试给予所有测试者，每一个项目对于判断分数的影响可能会变大。在这种情况下，检验CAT中的项目功能差异比起检验非自适应测验的项目功能差异，意义更为重大(Zwick, 2000)。

显然，CAT中的项目功能差异检验极其困难，充满了技

术层面和理解层面的挑战。然而尽管如此,检验仍然十分重要。首先,它对理解项目功能差异本身十分重要。其次,在CAT中,项目功能差异可能会比笔试中的明显。当存在非测试因素带来的功能异常时,项目功能差异更难被准确地检测和理解。

第4节 翻译的测试

在早期研究中，研究者即提出了使用项目功能差异方法研究翻译的、跨文化测试中的项目缺陷（Allalouf et al.，1999；Brislin，1970；Hulin、Drasgow & Komocar，1982）。不过，这一方法却很少被真正使用。然而，无论如何，对这一方法的使用可能是有效的。汉布尔顿（1993）提出了一个英文—瑞典语翻译测试的例子，在英文中，描述了一种具有“网状足”的鸟类，要求判断其适合的生存环境，而在瑞典语中，“网状足”被译为“游泳足”，因而暗示了正确答案。因此，对于使用两种语言的被测者而言，这个项目是不等同的。因而也无法进一步做出可信的跨文化比较。

此外，对于翻译的测试而言，没有哪一种项目功能差异检验方法是更合适的。当首要考量是文化差异时，我们需要考虑的是使用的项目功能差异方法是否适合具体的测试环境。

而且，由于越来越多的测试被翻译，有些测试已经发展出了具体的多语言版本。在这种情况下，研究项目功能差异十分必要。国际组织（如国际测试委员会）甚至列出了一些大纲，以帮助这一领域内的研究者（Hambleton，1994）。我们鼓励那些可能被用于跨文化环境测试的研究者进一步了

解项目功能差异。

另外,当越来越多的研究者关注如何评估种族文化背景不同的被测者时,项目功能差异也能被有效地用于测试和项目研发中(见 Gierl、Rogers & Klinger,1999)。

第12章

未来研究方向

正如我们在导言中所说,项目功能差异这一研究领域欣欣向荣。从前面的章节中我们也可以发现,项目功能差异的研究方法在不断发展。在前面几章,我们探讨了项目功能差异研究中一些特殊的方法。展望未来,我们发现以项目反应理论为基础的项目功能差异也在不断向前发展。项目反应理论方法最早由西森及同事提出(1993)。他们认为,现代的项目功能差异研究可以被看做多组项目反应理论的一个特例。

第1节 | 效度论题

项目功能差异研究的一个新方向是将项目看做效度问题的一个诊断性应用。这种方法被称为项目验证(Kane, 2006; Lissitz & Samuelsen, 2007)。这一视角直接支持了将效度看做一个评价论题的现代观点。这一观点由美国教育研究协会等于1999年在《教育和心理测试标准》中总结提出,并有支持者如梅西克(Messick, 1988、1989)。在这里,效度被看做在一个特定的评价情景中,评价从测试分数中获取的阐述的意义、有效性和恰当性。这种对效度的理解在心理测量专家内达成普遍共识,然而遗憾的是,另一些同样和测试打交道的人群(如一些老师、政策制定者或缺乏必要知识的测试研发者)并没有理解它。

效度论题建基于格利克森(Gulliksen, 1950)最早的对效度的讨论。他将效度理论化为测试项目和测试间的契合。这一观点与洛德(1980)对测试中项目功能的论述也是一致的。洛德认为,"一个项目的贡献,在某种复杂的意义上,依赖于测试中项目的选择"。

穆森(Muthen, 1988)将这种研究称为"微观层面的效度检验",并建议使用结构方程模型,将它作为验证性因子研究。

奥斯德兰(2006)指出,对测试的解读来源于项目的信息。因此,整个测试的效度假设各个项目的推论都是恰当并有意义的。从这里可以看出,当确定项目功能差异的效度论题时,仅仅检验被测者的项目反应是不够的。除此以外,专家对测试内容的检验、项目与内容的联系同样需要被包括进去。

第2节 | 原假设检验

正如前文所提到的，另一个和项目功能差异的效度论题密切相关的是一场项目的原假设检验。这也越来越受到学者的重视。这一方法完美地契合了将效度视为对测试分数的阐释这一观点。保罗·霍兰及其在教育测试机构的同事(Dorans & Holland, 1993; Holland & Wainer, 1993)提出了这一评价性论题。在 M-H 方法中，项目功能差异是通过使用传统的原假设检验方法进行评估的。原假设为不存在项目功能差异。也就是说，在这一框架下，原假设认为一个项目的难度与被测者的组别归属无关。相反，备择假设是项目特征和被测者组别归属间存在关系，即存在项目功能差异。

许多研究者不断指出教育和心理学研究中对于原假设检验方法的使用和滥用(见 Cohen、Kane & Crooks, 1998; Tryon, 2001)。他们的反对渐渐在项目功能差异研究中取得影响(Casabianca, 2008; Wells et al., 2008)。近期，如马库斯(Markus, 2001)、蒂龙(Tyron, 2001)和韦尔克(Wellek, 2002)概括的那样，越来越多的研究者使用两个单边检验进行相等性的检验。总的说来，这一方法通过涵盖两组共同定义“差异”的备择值以及一组定义“相等”的区间值，弥补了传统原假设检验方法的不足。

第3节 | 统计模型(HLM模型及其他)

在教育和心理学研究中,研究者越来越普遍使用统计模型进行研究。这一趋势也延伸到了项目功能差异领域的研究中。我们已经在依托于项目反应理论的方法中发现了这一点。对于项目功能差异的统计模型而言,对组间表现差别的考量被延伸,将一组项目视为更大范围内测试刺激的一个随机样本。在这种情况下,使用 logistic 混合模型,建模项目功能差异,将项目影响及其交互项看做随机的,而不是恒定的。类似地,组别也被视为从总体中较多的组别中随机选择的。

这种项目功能差异的建模方法在有些情况下可以帮助解释功能差异,甚至归因。可能的解释被当做协变量加入到模型中。因此,项目功能差异依赖于组别变量或者项目特征。需要注意的是,在这种情况下,解释并不需要是完美的,一些有用的信息仍可被获得。

在项目反应理论建模方法中,项目特征曲线的累积函数在组间是不同的。模型中展现了位置(即难度)和斜率(即区分度)。这一概念与项目反应理论模型契合得很好——项目反应理论模型也被视为一种 logistic 混合模型(见 Adams、Wilson & Wu, 1997; Bock & Aitkin, 1981)。

文·登·努尔盖特和伯克(Van den Noortgate & Boeck, 2005)提出,随机方法在一系列情境下都是可行的,存在三个优势。首先,logistic 模型适用于多种情况,并且对于具有相关知识储备的研究者而言极易理解。其次,从统计学上说,它们也更经济,因为只估计参数均值和方差,而不是所有的个体影响。最后,假设设想的解释,即便并不完整或完美,也可以被纳入为协变量。

等级线性模型(HLM)在项目功能差异研究中的使用虽然较新,但是有较大的发展潜力。在 HLM 中,将异常项目功能建模为一个项目反应的双层构成。能力被建模为一个随机影响,而非像是在项目反应理论模型中那样,被用以限制项目参数。在项目功能差异研究中使用 HLM 时,模型的随机部分(如随机系数)被看做普通的误差测量,而不被强调。这一方法虽然较新,但近来在不断发展。卡马塔(Kamata, 2001)在这一领域做着值得一提的研究。

第4节 | 等同性检验

在韦尔斯等人(Wells et al.，2008)近期的一篇文章中，在依据洛德卡方检验的项目功能差异项目反应理论模型中，使用了等同性检验的方法。在一个较大的样本下，他们展示了等同性检验在项目功能差异框架下的用处，并找到了相对稳健的降低测试长度的方法。在此项研究的延伸中，卡萨比安卡(Casabianca，2008)发展出一套方法，她将其称为“等同项目功能”，这一方法使用两个单边检验(Tyron，2001；Wellek，2002)。

卡萨比安卡使用单参数模型的虚拟数据，通过变化样本大小和项目功能差异的程度，来检验对于M-H估计“等同项目功能”方法的稳健程度。她的结论是，这一新方法，虽然有一定局限，但对于项目功能差异研究是可行的。同时其优势在于，为了解一个项目在不同组别间的被测者中是否一致提供了信息。

第 5 节 | DFIT 与 CDIF 检验

另一个新的方向是由拉古(1988)和同事发展出的 DFIT 方法。这一方法也是评估项目功能差异的一个有效的、适用性广的方法。这一方法可以用以检验二分类和多分类测试项目,也适用于一维以及多维测试。DFIT 来源于项目反应理论的参数估计,因而需要一个较大的样本。除了在项目层面上进行标准的项目功能差异估计外,这一方法还能够估计补偿性(CDIF)以及非补偿性项目功能差异(NCDIF)。更有用的是,这一方法在评估 DTF(测试功能差异)方面也具有优势。

补偿性项目功能差异概念(由 CDIF 指数代表)的优势在于,允许研究者评估去除一个特定的测试项目对于测试功能差异的估计的影响。在这一框架下,心理测量专家和测试研发者便有可能制定出对分数的差异影响最小的测试。

拉古及其同事(1995)最早根据卡方统计量设计了 DFIT 的显著性检验,其核心目标在于制定出可以检验 CDIF,以及最终检验 DTF 的统计方法。这些卡方检验随着时间推移被进一步完善。而最近又有研究者提出了新的项目参数复制方法(Oshima、Raju & Nanda, 2006)。研究者需要重视的是,这些新的方法在最新版的 DFIT 软件中已被使用(见 Raju

et al.，2005)。欲了解此方面的更多信息,可参阅奥谢玛等人(Oshima et al.，2006)的著作。

可见,从实践操作和新方法研发的角度讲,越来越多的方法指向越来越多的项目功能差异研究。

第13章

总　结

在过去的几十年间,对于组间表现区别的研究发生了巨大的改变。项目功能差异的概念比以往更广,涵盖了公平性、偏见、负面影响等许多重要议题。用来研究系统的组间表现差异的方法也进一步发展。新方法被研究出来,不合时宜的旧方法被抛弃。从实践操作的角度,我们描述和解释了一系列对于项目功能差异检验普遍的和重要的统计方法。我们从始至终的目标在于,与《测试项目偏差》相比,进一步介绍这一领域的发展和变化。这是一个广阔的领域,我们的目标是介绍其中最新的发展,尽管我们也简略介绍了它的历史。在本书中,我们并不试图追求完备,我们所说的也并非绝对确定。因此,我们大量引用了他人的研究成果。我们共同加深了对项目功能差异研究理论和方法的认识与理解。我们希望研究者、学者、学生与测试研发者在研究项目功能差异时,本书能有所助益。

参考文献

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 674—691.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to error in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47—76.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185—198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland & W. Wainer(Eds.), *Differential item functioning*(pp. 3—24). Baltimore: Johns Hopkins University Press.

Baker, F. B. (2001). *The basics of item response theory*. Retrieved May 5, 2009, from http://echo.edres.org:8080/irt/baker/

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting item bias*. Baltimore: Johns Hopkins University Press.

Birnbaum, A. (1958). *On the estimation of mental abilities*(Series Report No 115. Project 7755-23). Randolf Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. In F. M. Lord & M. R. Novick(Eds.), *Statistical theories of mental test scores*(pp. 395—479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443—445.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113—141.

Bolt, D. M. (2005). Limited and full-information IRT estimation. In A. Maydeu-Olivares & J. McArdie (Eds.), *Contemporary psychometrics* (pp. 27—71). Hillsdale, NJ: Lawrence Erlbaum.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185—216.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 220—256), Westport, CT: American Council on Education.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Casabianca, J. M. (2008). *Equivalence testing for differential item function detection*. Unpublished draft of master of arts thesis, Fordham University, New York.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31—44.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115—124.

Clifford, H. W. (1982). Simpson's Paradox in real life. *The American Statistician*, *36*, 46—48.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1998). A generalized examinee-centered method for setting standards on achievement tests. *Applied Psychological Measurement*, *12*, 343—366.

Cole, N., & Moss, P. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education/Macmillan.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 25). Hillsdale, NJ: Lawrence Erlbaum.

Dodeen, H., & Johanson, G. A. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment and Evaluation in Higher Education*, *28*, 129—134.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35—66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (Research Report No. 83—9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the SAT. *Journal of Educational Measurement*, *23*, 355—368.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19—29.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International. Retrieved May 5, 2009, from www. ssicentral. com.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fildago, A. M., Ferreres, D., & Muniz, J. (2005). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *Journal of Experimental Education*, *73*, 23—39.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278—295.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*, *23*(4), 309—326.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*, 315—332.

Gelin, M. N., Carelton, B. C., Smith, A. A., & Zumbo, B. D. (2004). The dimensionality and gender differential item functioning of the mini asthma quality of life questionnaire (MINIAQLQ). *Social Indicators Research*, *68*, 91—105.

Gierl, M., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education*, *17*(3), 241—264.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). *Using statistical and judgmental reviews to identify and interpret translation DIF*. Paper presented at the National Council on Measurement in Education, New Orleans, LA.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer et al. (Eds.), *Measurement and prediction: The American soldier* (Vol. IV). New York: Wiley.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *Journal of Psychological Assessment*, *9*, 57—68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229—244.

Hambleton, R. K., & Bollward, J. (1990). *Factors affecting the stability of Mantel-Haenszel item bias statistics*. Amherst: University of Massachusetts.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijoff.

Hambleton, R. K., Swaminathan H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P. W. (1985). *On the study of differential item performance without IRT*. Paper presented at the Proceedings of the Military Testing Association.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129—145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, *67*, 818—825.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79—93.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17—64). Westport, CT: Praeger.

Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement*, *45*(3), 271—285.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531—578). Westport, CT: American Council on Education/Praeger.

Kwak, N., Davenport, E. D., & Davison, M. L. (1998, April). *A comparative study of observed score approaches and purification procedures for detecting differential item functioning*. Paper presented at the National Council on Measurement in Education, Denver, CO.

Lane, S., & Stone, C. A. (2006). Performance assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 422—468). Westport, CT: American Council on Education/Praeger.

Lange, R., Thalbourne, M. A., Houran, J., & Lester, D. (2002). Depressive response sets due to gender and culture-based differential item functioning. *Personality and Individual Differences*, *33*, 937—954.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437—448.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719—748.

Markus, K. A. (2001). The converse of inequality argument against tests of statistical significance. *Psychological Methods*, *6*, 147—160.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105—118.

Messick, S. (1988). The once and future issues of validity: Assessing the

meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33—46). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn(Ed.), *Educational measurement* (3rd ed., pp. 13—105). New York: American Council on Education/Macmillan.

Mills, C. N., Portenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-based testing*. Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E., & Everson, H. T. (1993). Methodological review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297—334.

Muraki, E. (1992). *RESGEN*(No. RR-92-7). Princeton, NJ: Educational Testing Service.

Muraki, E., & Bock, D. (2003). *PARSCALE: IRT based test scoring and item analysis for graded response items and rating scales* (Version 4. 1). Lincolnwood, IL: Scientific Software International.

Muthen, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun(Eds.), *Test validity*. Mahwah, NJ: Lawrence Erlbaum.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests(DFIT) framework. *Journal of Educational Measurement*, *43*, 1—17.

Osterlind, S. J. (1983). *Test item bias* (Vol. 30). Beverly Hills, CA: Sage.

Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Prentice Hall.

Osterlind, S. J., Sheng, Y., Wang, Z., Beaujean, A. A., & Nagel, T. (2008). *Technical manual: College Basic Subjects Examination*. Columbia: University of Missouri-Columbia.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay(Eds.), *Handbook of statistics* (Vol. 26, pp. 125—167). New York: Elsevier.

Penfield, R. D., & Gattamorta, K. (2009). An NCME instructional module using differential step functioning to refine the analyis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28* (1),

38—49.

Penfield, R. D. , & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement Issues and Practice*, *19*(3), 5—15.

Potenza, M. T. , & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*, 23—37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495—502.

Raju, N. S. , Oshima, T. C. , & Wolach, A. (2005). Differential functioning of items and tests(DFIT): Dichotomous and polytomous[Computer program]. Chicago: Illinois Institute of Technology.

Raju, N. S. , van der Linden, W. J. , & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 153—168.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Pedagogiske Institute. (Reprinted in 1980, Chicago: University of Chicago Press)

Roussos, L. A. , & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355—371.

Rudner, L. , & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, *7*(26). Retrieved May 5, 2009, from http://ericae.net/pare/getvn.asp?v=7&n=26.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 67—84). New York: Springer.

Shealy, R. , & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159—194.

Shepard, L. , Camilli, G. , & Averil, M. (1981). Comparison of procedures for detecting testitem bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*, 317—375.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society B*, *13*, 238—241.

SPSS. (2008). SPSS 16.0 for Windows(Version 16.0). Chicago: Author.

Stout, W. (1995). SIBTEST: Differential items/bundle functioning. St. Paul, MN: Assessment Systems Corporation.

Thissen, D., Chen, W., & Bock, D. (2003). MULTILOG: multiple category item analysis and test scoring using item response theory(Version 7). Lincolnwood, IL: Scientific Software International.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun(Eds.), *Test validity*(pp. 147—170). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67—113). Hillsdale, NJ: Lawrence Erlbaum.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: An integrated alternative method of conducing null hypothesis statistical tests. *Psychological Methods*, *6*(4), 371—386.

van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443—464.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston: Kluwer.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2008). Comparability of computerbased and paper-and-pencil testing in K-12 reading assessments. *Educational and Psychological Measurement*, *68* (1), 5—24.

Wellek, S. (2002). *Testing statistical hypotheses of equivalence*. Boca Raton, FL: Chapman & Hall.

Wells, C. S., Wollack, J. A., & Serlin, R. C. (2008, April). *An equivalency test for model DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement*, *59* (4), 910—927.

Wilder, G., & Powell, K. (1989). Sex differences in test performance: A

survey of the literature(Report No. RR 89-4). Princeton, NJ: Educational Testing Service.

Williams, B. (1978). *A sampler on sampling*. New York: Wiley.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2003). TESTFACT: Test scoring, item statistics, and item factor analysis (Version 4. 0). Lincolnwood, IL: Scientific Software International.

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359—376). Mahwah, NJ: Lawrence Erlbaum.

Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, National Defense Headquarters.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 275—321). Boston: Kluwer.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential items functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233—251.

译名对照表

ability	能力
adverse impact	负面影响
Analysis of Variance(ANOVA)	方差分析
bias	偏见,偏差
Chi-Square contingency table	卡方列联表
Civil Rights Act	《民权法案》
classical item analysis	经典项目分析
College Basic Academic Subjects Examination	大学基本学科考试
Computer-Adaptive Testing(CAT)	计算机自适性测试
computer-based testing	机考
differential Item Functioning	项目功能差异
error in measurement	测量误差
fairness	公平性
formal definition	正式定义
item response theory	项目反应理论
logistic regression	logistic 回归
nonparametric methods	非参数方法
nonuniform	不统一的,非统一的
null hypothesis testing	假设检验
statistical modeling	统计建模
Dorans' standardization	Dorans 标准化
Hierarchal Linear Modeling (HLM)	等级线性模型
Item Characteristic Curve(ICC)	项目特征曲线
item parameter	项目参数
one-parameter model	单参数模型
two-parameter model	双参数模型
three-parameter model	三参数模型
parameter estimates	参数估计
item	项目
item impact	项目影响
likelihood ration test	似然比检验

Mantel-Haenszel Test	Mantel-Haenszel(M-H)检验
odds ratio	比值比
polytomously scored item	多元评分项目
purification of data	数据清洁
Simpson's paradox	Simpson 悖论
statistical bias	统计偏差
validity	效度

Differential Item Functioning, 2nd Edition

Copyright © 2009 by SAGE Publications, Inc.
All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.
This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2013.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。
上海市版权局著作权合同登记号:图字 09-2009-552